The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Number 4 December 1950 Volume 6

Material for *Biometrics* should be addressed to the Chairman of the Editorial Board, Institute of Statistics, North Carolina State College, Raleigh, N. C.; and material for Queries should go to "Queries", Statistical Laboratory, Iowa State College, Ames, Iowa, or to any member of the committee.

Articles to be considered for publication in *Biometrics* should be submitted in triplicate.

The Biometric Society is an international society devoted to the mathematical and statistical aspects of biology and welcomes to membership biologists, mathematicians, statisticians and others who are interested in its objectives. Through its six regional organizations the Society sponsors regional and local meetings. National secretaries serve the interests of members in Italy, Denmark and the Netherlands and there are many members "at large". Dues in the Society for 1951 are as follows: Full membership including subscription to *Biometrics* is $7.00. Members of the Biometrics Section of the American Statistical Association who subscribe to the journal through that organization may become members of The Biometric Society on the payment of $3.00 annual dues.

Annual subscription rates to non-members are as follows: For American Statistical Association Members, $4.00; for subscribers, non-members of either American Statistical Association or The Biometric Society, $7.00. Subscription and application should be sent to The Biometric Society, 52 Hillhouse Ave. New Haven, Connecticut, U.S.A.

Entered as second-class matter at the Post Office at New Haven, Conn., under the Act of March 3, 1879. Additional entry at Richmond, Va. and Raleigh, North Carolina. Business Office, 52 Hillhouse Ave., New Haven, Conn. *Biometrics* is published quarterly—in March, June, September and December.

GENE FREQUENCIES IN A CLINE DETERMINED BY SELECTION AND DIFFUSION

R. A. Fisher

University of Cambridge

1. *The ecological problem.*

In 1937 (1) the author studied the distribution of the gene ratio in the simple case of an advantageous gene advancing along a linear habitat under a constant selective advantage. In Nature situations must be more complex than in this simple model; in particular it must often occur that the selective advantage itself varies with position. The interesting case arises in which a gene enjoys a selective advantage in one part of a species' range, while in the remainder it is at a selective disadvantage. On the boundary between these regions selection is neutral between two allelomorphic genes.

Cases can be observed in practice in which there is a gradient, or cline, in the frequencies of the genotypes determined by a single factor. Generally such cases will be complicated by inequalities of topography, and by consequent irregularities in the population density, and in the gradient of selective advantage. It is to be expected also that the boundary will in general neither be straight (i.e. a great circle of the earth's surface), nor constant in position under the conditions prevailing in different years. Any model worth discussing from a theoretical standpoint will therefore be a drastically simplified one, playing the part of a basis for comparisons by which the real complexities of each situation may be critically demonstrated.

The purely genetical complication of non-recognition of genotype, due to dominance, is also ignored. This is chiefly because I should regard the occurrence of true dominance in such cases as a danger-signal suggesting the very different genetic situation of a balanced polymorphism; partly because if on more careful examination it is found that the heterozygote is recognisable this will greatly increase the value of the observations.

The problem chosen for discussion, as an extension of my work of 1937, thus differs materially from that considered by Haldane in 1948 (2). Haldane discusses the effects of a discontinuous selective intensity acting on a gene ratio obscured by dominance.

The available data, in the common case, will consist of gene frequencies observed at chosen centres of collection. From such data the primary need is to determine the neutral, or 50%, line, and the distances

TABLE I

FUNDAMENTAL TABLE, GIVING THE SOLUTION OF THE DIFFERENTIAL EQUATION (1)

x	q	x	q	x	q
.00	.5000 0000	1.00	.1004 2286	2.00	.0077 7553
.02	.4895 1717	1.02	.0962 0743	2.02	.0073 3125
.04	.4790 4235	1.04	.0921 3390	2.04	.0069 1050
.06	.4685 8348	1.06	.0881 9955	2.06	.0065 1215
.08	.4581 4852	1.08	.0844 0159	2.08	.0061 3513
.10	.4477 4533	1.10	.0807 3716	2.10	.0057 7840
.12	.4373 8169	1.12	.0772 0335	2.12	.0054 4098
.14	.4270 6529	1.14	.0737 9721	2.14	.0051 2192
.16	.4168 0369	1.16	.0705 1573	2.16	.0048 2031
.18	.4066 0430	1.18	.0673 5591	2.18	.0045 3528
.20	.3964 7438	1.20	.0643 1468	2.20	.0042 6600
.22	.3864 2102	1.22	.0613 8899	2.22	.0040 1168
.24	.3764 5110	1.24	.0585 7578	2.24	.0037 7155
.26	.3665 7130	1.26	.0558 7198	2.26	.0035 4489
.28	.3567 8807	1.28	.0532 7452	2.28	.0033 3101
.30	.3471 0763	1.30	.0507 8036	2.30	.0031 2924
.32	.3375 3596	1.32	.0483 8646	2.32	.0029 3895
.34	.3280 7874	1.34	.0460 8981	2.34	.0027 5954
.36	.3187 4142	1.36	.0438 8742	2.36	.0025 9044
.38	.3095 2916	1.38	.0417 7635	2.38	.0024 3109
.40	.3004 4681	1.40	.0397 5367	2.40	.0022 8099
.42	.2914 9895	1.42	.0378 1649	2.42	.0021 3962
.44	.2826 8985	1.44	.0359 6200	2.44	.0020 0652
.46	.2740 2349	1.46	.0341 8738	2.46	.0018 8124
.48	.2655 0351	1.48	.0324 8990	2.48	.0017 6336
.50	.2571 3327	1.50	.0308 6686	2.50	.0016 5246
.52	.2489 1582	1.52	.0293 1561	2.52	.0015 4816
.54	.2408 5390	1.54	.0278 3358	2.54	.0014 5010
.56	.2329 4992	1.56	.0264 1823	2.56	.0013 5792
.58	.2252 0602	1.58	.0250 6707	2.58	.0012 7130
.60	.2176 2403	1.60	.0237 7771	2.60	.0011 8992
.62	.2102 0546	1.62	.0225 4777	2.62	.0011 1348
.64	.2029 5155	1.64	.0213 7496	2.64	.0010 4171
.66	.1958 6327	1.66	.0202 5705	2.66	.0009 7434
.68	.1889 4129	1.68	.0191 9186	2.68	.0009 1111
.70	.1821 8602	1.70	.0181 7727	2.70	.0008 5178
.72	.1755 9759	1.72	.0172 1122	2.72	.0007 9613
.74	.1691 7592	1.74	.0162 9174	2.74	.0007 4395
.76	.1629 2064	1.76	.0154 1687	2.76	.0006 9502
.78	.1568 3117	1.78	.0145 8475	2.78	.0006 4916
.80	.1509 0672	1.80	.0137 9358	2.80	.0006 0619
.82	.1451 4627	1.82	.0130 4158	2.82	.0005 6594
.84	.1395 4859	1.84	.0123 2707	2.84	.0005 2823
.86	.1341 1227	1.86	.0116 4841	2.86	.0004 9293
.88	.1288 3573	1.88	.0110 0401	2.88	.0004 5988
.90	.1237 1721	1.90	.0103 9235	2.90	.0004 2895
.92	.1187 5479	1.92	.0098 1197	2.92	.0004 0001
.94	.1139 4640	1.94	.0092 6143	2.94	.0003 7294
.96	.1092 8985	1.96	.0087 3938	2.96	.0003 4763
.98	.1047 8282	1.98	.0082 4450	2.98	.0003 2396

TABLE I—*Continued*

FUNDAMENTAL TABLE, GIVING THE SOLUTION OF THE DIFFERENTIAL EQUATION (1)

x	q	x	q	x	q
3.00	.0003 0183	4.00	.0000 0670	5.00	.0000 0009
3.02	.0002 8115	4.02	.0000 0618	5.02	.0000 0008
3.04	.0002 6183	4.04	.0000 0570	5.04	.0000 0008
3.06	.0002 4379	4.06	.0000 0525	5.06	.0000 0007
3.08	.0002 2694	4.08	.0000 0484	5.08	.0000 0006
3.10	.0002 1121	4.10	.0000 0446	5.10	.0000 0006
3.12	.0001 9652	4.12	.0000 0410	5.12	.0000 0005
3.14	.0001 8282	4.14	.0000 0378	5.14	.0000 0005
3.16	.0001 7003	4.16	.0000 0348	5.16	.0000 0004
3.18	.0001 5811	4.18	.0000 0320	5.18	.0000 0004
3.20	.0001 4699	4.20	.0000 0295	5.20	.0000 0004
3.22	.0001 3662	4.22	.0000 0271	5.22	.0000 0003
3.24	.0001 2696	4.24	.0000 0250	5.24	.0000 0003
3.26	.0001 1795	4.26	.0000 0230	5.26	.0000 0003
3.28	.0001 0956	4.28	.0000 0211	5.28	.0000 0003
3.30	.0001 0175	4.30	.0000 0194	5.30	.0000 0002
3.32	.0000 9447	4.32	.0000 0178	5.32	.0000 0002
3.34	.0000 8769	4.34	.0000 0164	5.34	.0000 0002
3.36	.0000 8139	4.36	.0000 0151	5.36	.0000 0002
3.38	.0000 7552	4.38	.0000 0138	5.38	.0000 0002
3.40	.0000 7006	4.40	.0000 0127	5.40	.0000 0001
3.42	.0000 6498	4.42	.0000 0117	5.42	.0000 0001
3.44	.0000 6026	4.44	.0000 0107	5.44	.0000 0001
3.46	.0000 5586	4.46	.0000 0098	5.46	.0000 0001
3.48	.0000 5178	4.48	.0000 0090	5.48	.0000 0001
3.50	.0000 4799	4.50	.0000 0083	5.50	.0000 0001
3.52	.0000 4446	4.52	.0000 0076	5.52	.0000 0001
3.54	.0000 4119	4.54	.0000 0070	5.54	.0000 0001
3.56	.0000 3815	4.56	.0000 0064	5.56	.0000 0001
3.58	.0000 3532	4.58	.0000 0059	5.58	.0000 0001
3.60	.0000 3270	4.60	.0000 0054	5.60	.0000 0001
3.62	.0000 3027	4.62	.0000 0049	5.62	.0000 0001
3.64	.0000 2801	4.64	.0000 0045	5.64	.0000 0000
3.66	.0000 2591	4.66	.0000 0041	5.66	.0000 0000
3.68	.0000 2397	4.68	.0000 0038	5.68	.0000 0000
3.70	.0000 2217	4.70	.0000 0035	5.70	.0000 0000
3.72	.0000 2050	4.72	.0000 0032		
3.74	.0000 1895	4.74	.0000 0029		
3.76	.0000 1752	4.76	.0000 0027		
3.78	.0000 1619	4.78	.0000 0024		
3.80	.0000 1495	4.80	.0000 0022		
3.82	.0000 1381	4.82	.0000 0020		
3.84	.0000 1276	4.84	.0000 0019		
3.86	.0000 1178	4.86	.0000 0017		
3.88	.0000 1087	4.88	.0000 0016		
3.90	.0000 1004	4.90	.0000 0014		
3.92	.0000 0926	4.92	.0000 0013		
3.94	.0000 0855	4.94	.0000 0012		
3.96	.0000 0788	4.96	.0000 0011		
3.98	.0000 0727	4.98	.0000 0010		

from this at which other percentages are to be expected. The scale of these distances is an important observational feature, depending on the diffusional mobility of the species, and on the intensity of the selective gradient.

2. *Mathematical formulation.*

The frequency p of a gene depends only on the coordinate x, as in a linear habitat.

In the case in which the gene with frequency p has a selective advantage i (defined as rate of change of logarithmic gene ratio, (3) pp. 70–72)) proportional to x, and x varies without limit in both directions, the rate of increase due to selection is

$$\frac{dp}{dt} = pqi = pq\ g\ x,$$

where g is the gradient of selective advantage, and $q = 1 - p$.

Suppose also a uniform population density with a diffusion constant k such that the rate of increase in gene frequency (p) at any point is

$$k\frac{d^2p}{dx^2} = -k\frac{d^2q}{dx^2}$$

due to diffusion. In 1937 I discussed the limitations and justification of the analogy with physical diffusion.

Then the gene ratio will adjust itself so as to tend to satisfy the equation of equilibrium,

$$k\frac{d^2q}{dx^2} = pq\ g\ x;$$

we may choose the unit of length in which the position x is measured, so that

$$g = 4k,$$

and the relation between q and x is then given by the equation

$$\left.\begin{aligned} &\frac{d^2q}{dx^2} = 4\ x\ pq, \\ &\text{with the boundary conditions} \\ &x = 0 \qquad q = \tfrac{1}{2} \\ &x = \infty \qquad q = 0 \end{aligned}\right\} \tag{1}$$

TABLE II

STANDARDIZED DEVIATES (LEGITS) FOR GIVEN GENE PERCENTAGES

Gene percentage	Deviate	Gene percentage	Deviate	Gene percentage	Deviate
50	.00000	20.0	.64827	5.0	1.30643
49	.01908	19.5	.66247	4.8	1.32331
48	.03817	19.0	.67691	4.6	1.34080
47	.05729	18.5	.69161	4.4	1.35896
46	.07645	18.0	.70658	4.2	1.37784
45	.09566	17.5	.72184	4.0	1.39752
44	.11493	17.0	.73740	3.8	1.41807
43	.13429	16.5	.75329	3.6	1.43958
42	.15375	16.0	.76952	3.4	1.46216
41	.17332	15.5	.78612	3.2	1.48594
40	.19302	15.0	.80311	3.0	1.51106
39	.21286	14.5	.82052	2.8	1.53771
38	.23286	14.0	.83837	2.6	1.56609
37	.25304	13.5	.85669	2.4	1.59648
36	.27341	13.0	.87553	2.2	1.62922
35	.29400	12.5	.89492	2.0	1.66474
34	.31483	12.0	.91492	1.8	1.70360
33	.33592	11.5	.93556	1.6	1.74656
32	.35729	11.0	.95691	1.4	1.79468
31	.37897	10.5	.97902	1.2	1.84951
30	.40099	10.0	1.00198	1.0	1.91340
29	.42338	9.5	1.02586	0.9	1.94988
28	.44617	9.0	1.05076	0.8	1.99029
27	.46940	8.5	1.07680	0.7	2.03565
26	.49311	8.0	1.10411	0.6	2.08745
25	.51734	7.5	1.13285	0.5	2.14795
24	.54214	7.0	1.16321	0.4	2.22095
23	.56757	6.5	1.19543	0.3	2.31345
22	.59369	6.0	1.22978	0.2	2.44101
21	.62056	5.5	1.26662	0.1	2.65223

Values of q to eight decimal places, from $x = 0$, by intervals of .02, to extinction, are given in Table I. I owe this tabulation to Dr. M. V. Wilkes and Mr. D. J. Wheeler, operating the EDSAC electronic computer. The last decimal place may be in error by 3 or 4 units. From this primary Table I have calculated Table II, giving the deviation x

TABLE III

APPARATUS FOR FINAL FITTING

Provisional Legit	Working Legit		Weighting Coeff.	Provisional Legit	Working Legit		Weighting Coeff.
	Max.	Min.			Max.	Min.	
.00	.9538	−.9538	1.0992	1.50	1.8891	−10.7156	.2104
.10	.9623	−.9636	1.0903	1.60	1.9776	−13.905	.1708
.20	.9857	−.9960	1.0643	1.70	2.0673	−18.135	.1373
.30	1.0211	−1.0564	1.0224	1.80	2.1577	−23.773	.1093
.40	1.0665	−1.1518	.9669	1.90	2.2489	−31.321	.0863
.50	1.1200	−1.2913	.9004	2.00	2.3407	−41.478	.0675
.60	1.1803	−1.4862	.8260	2.10	2.4331	−55.218	.0524
.70	1.2461	−1.7515	.7469	2.20	2.5261	−73.906	.0403
.80	1.3166	−2.1067	.6659	2.30	2.6194	−99.459	.0308
.90	1.3909	−2.5772	.5858	2.40	2.7132	−134.599	.0233
1.00	1.4685	−3.1965	.5087	2.50	2.8073	−183.17	.0175
1.10	1.5487	−4.0090	.4362	2.60	2.9018	−250.70	.0131
1.20	1.6312	−5.0736	.3697	2.70	2.9965	−345.11	.0097
1.30	1.7156	−6.4693	.3097	2.80	3.0916	−477.90	.0071
1.40	1.8017	−8.3021	.2566	2.90	3.1868	−665.39	.0052
1.50	1.8891	−10.7156	.2104	3.00	3.2822	−931.73	.0038

corresponding with 90 chosen probabilities q, and Table III, supplying the computational apparatus analogous to that used in probit analysis of mortality data in toxicology.

To appreciate this analogy q may be regarded as the probability of a variate exceeding x in a certain symmetrical distribution. x is thus a transformation of q which, under the hypothesis, is linear with—and subject to choice of units as explained above is equal to—the distance from the neutral line.

For the normal distribution with unit variance this probability tends to zero in such a way that

$$\log 1/q \div \tfrac{1}{2}x^2 \to 1$$

as x is increased. For the distribution for which

$$2x = \log p - \log q,$$

$$\frac{dp}{dx} = \frac{1}{2}\operatorname{sech}^2 x,$$

we have

$$\log 1/q \div 2x \to 1.$$

The case with which we are concerned falls between these two, for $\log 1/q$ tends at the limit to equality with $\frac{4}{3} x^{3/2}$.

The relation between the three forms may also be shown from their properties at or near the median. Here we may consider

$$-\frac{d^3p}{dx^3} \div \left(\frac{dp}{dx}\right)^3,$$

which is dimensionless, and therefore a pure measure of form. For this ratio we find

		ratio	
Normal transformation	Probit	2π	6.2832
Logarithmic transformation	Logit	2^3	8.0000
Selection-diffusion transformation	Legit	$(1.90764)^3$	6.9421

Since it has been found convenient to distinguish the deviates in the two first cases as Probits and Logits respectively, a similar term such as *Legit* may be found convenient for the standardised deviate of the distribution obtained above from a uniform cline of selection.

The dimensional nature of the quantities k and g involved in the provisional identity

$$g = 4k$$

is determined by

$$k \propto L^2T^{-1}$$

$$g \propto L^{-1}T^{-1},$$

hence

$$\frac{4k}{g} \propto L^3.$$

In setting this quantity equal to unity, therefore, we are merely adopting a unit of length appropriate to our problem, and this unit of length may now be defined as

$$a^3 = \frac{4k}{g}.$$

a will then be the distance separating the line on which q is 10.04 . . .% from the line of 50%, and that in turn from the line having 89.96 . . .% of the chosen gene.

3. *Fitting the data.*

For any observed gene-frequency the corresponding deviate can be read off from Table II. Since the primary data allow of it, five places of decimals have been given. It should, I think, only be used to three places. Gene frequencies (q) greater than 50% have corresponding negative deviates. Each empirical percentage has been based on twice as many genes as organisms have been classified; these double numbers should be multiplied by the weighting factors of Table III corresponding with the deviate obtained. No high precision of weighting is usually called for, since the data will probably contain some observations with 0 or 100%, and these cannot be included at the first stage. In all other cases the observed sample supplies a deviate, an appropriate weight, and two geographical coordinates λ and μ determining its position.

The first step is then to find a weighted regression equation,

$$X = b_0 + b_1\lambda + b_2\mu,$$

in which the coefficients b_0 , b_1 and b_2 have been adjusted so as to minimise the weighted sum of squares

$$S\{w(x - X)^2\}.$$

The (first) estimated position of the neutral line is then given by the equation,

$$b_0 + b_1\lambda + b_2\mu = 0;$$

and its distance from the parallel lines

$$b_0 + b_1\lambda + b_2\mu = \pm 1$$

provides an estimate of the scaling distance a.

No great labour need be expended on this stage of the work, the purpose of which is to provide provisional deviates X, with which to proceed. Usually indeed somewhat simplified values of b_0 , b_1 and b_2 are next substituted for those actually obtained, using values judged to be sufficiently near to those to be finally obtained.

For what will probably be the final fitting these provisional values, X, corresponding to the position of each sample observed, will be used to enter Table III. If 0% have been observed for q the working Legit has its maximum value as tabulated. Linear interpolation in x is generally sufficient. If 100%, the minimum is used, and for intermediate percentages the difference between these is divided proportionally. It is usually convenient at this stage to multiply the minimum by the actual number of genes recorded, the maximum by the number of allelomorphic

genes, and to divide the total by twice the number of organisms observed, before interpolation for the provisional value x required. The weighting coefficients corresponding with the same provisional values, multiplied by twice the observed number of organisms, give the working weights.

We now have working values for deviates, and weights, sufficiently accurate for a final determination of the regression of Legit on geographical position.

The values tabulated in Table III, in accordance with the general application of the Method of Maximum Likelihood are

$$\text{Maximum} \qquad X + Q\frac{\partial X}{\partial P}$$

$$\text{Minimum} \qquad X - P\frac{\partial X}{\partial P}$$

$$\text{Weighting coefficient} \qquad \frac{1}{PQ}\left(\frac{\partial P}{\partial X}\right)^2,$$

where P, Q stand for the probabilities corresponding to the provisional value X.

Using the analysis in this form, if s samples have been observed, with a sufficient number of both allelomorphic genes in each, the minimised value of $S\{w(x - X)^2\}$ may be taken as χ^2 with $(s - 3)$ degrees of freedom to test the goodness of fit of the hypothesis. In all respects the analogy with probit analysis is close.

The determination of the distance a provides the ratio of the diffusion coefficient k to the selective gradient. Since diffusive activity can sometimes be measured independently, the way is opened for the discussion, subject of course to complicating factors, of the selective gradient.

REFERENCES

(1) R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics, vii*, 355–369, 1937.

(2) J. B. S. Haldane. The theory of a cline. *Journal of Genetics, 48*, 277–284, 1948.

(3) R. A. Fisher. *The Genetical Theory of Natural Selection*. 70-72, 1930.

PROBLEMS IN THE ANALYSIS OF GROWTH AND WEAR CURVES

G. E. P. Box

Imperial Chemical Industries Limited
Dyestuffs Division Headquarters
Blackley, Manchester, England

1. INTRODUCTION

IN A NUMBER of different fields of application, the problem arises of comparing sets of growth and wear curves. The object of this paper is to describe methods of analysis which have been found useful, to point out what assumptions are being made, and show how these assumptions can themselves be put to the test.

1.1 Examples of wear and growth curves.

Much research is carried on by technologists with the object of improving the abrasion resistance of materials, and various machines have been devised to test this property. In most of these, specimens of materials to be compared are rubbed against a standard abrasive under standard conditions and the loss in weight or decrease in thickness of each of the specimens is noted at suitable intervals. For example, Fig. I shows the weight loss curves of 24 specimens of coated fabrics tested in the Martindale wear tester. The experiment was arranged in the form of a $2 \times 2 \times 3$ factorial design replicated twice, two different fillers F_1 and F_2 being tried in three different proportions Q_1, Q_2, and Q_3 with and without a surface treatment T, and weight losses were recorded after 1000, 2000, and 3000 revolutions of the machine.

Further examples of this type of test arose in road trial assessments of materials for tire treads (see for example Buist et al., 1950). In one of these investigations, for instance, a test vehicle was run with each of its tires constructed in 3 segments: 4 compounds A, B, C, and D were tested on the four tires of the car in a balanced incomplete block design as follows:

Tire 1	Tire 2	Tire 3	Tire 4
$B\ C\ D$	$A\ C\ D$	$A\ B\ D$	$A\ B\ C$

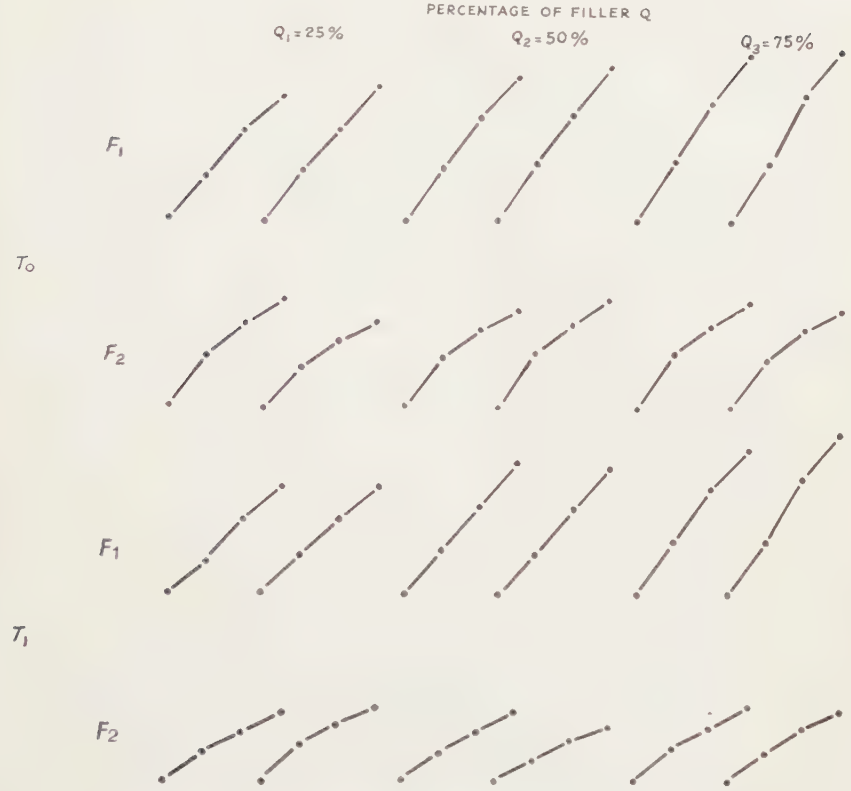


FIGURE I. WEIGHT LOSS CURVES FOR COATED FABRICS

The wear of the compounds was measured by the decrease in tread depth which was observed for each segment of each tire at a number of mileages. Thus for each compound in each tire, there again resulted, not a single result, but a set of results from each of which a wear curve could be plotted.

In biological investigations the growth of an animal or part of an animal is often the subject of study; Fig. II, for example, shows growth curves for 27 rats kept in separate cages. The rats were divided at random into 3 groups containing 10, 7, and 10 rats respectively, (the second group contains fewer rats than the other two, due to an accident at the beginning of the experiment). The first group were kept as a control, the second group had thyroxin, and the third group thiouracil added to their drinking water.

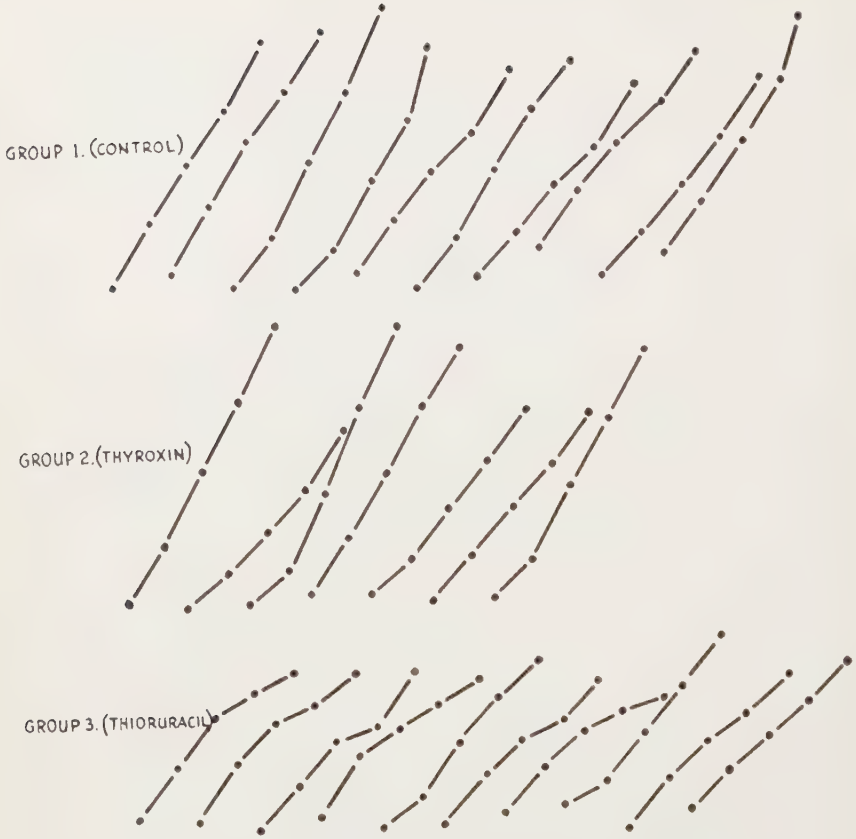


FIGURE II. GROWTH CURVES FOR 27 RATS

In all the examples, it will be seen that we are concerned with experiments which may be of a simple or complex character; each observation, however, consists not of a single value, but of a set of values recorded at intervals of time, from which curves can be plotted.

2. A SIMPLE ANALYSIS

We begin by considering the wear data plotted in Fig. I for coated fabrics prepared in a number of different ways. With data of this kind it has been found to be of value to consider not the wear, after say 1000, 2000, and 3000 revolutions of the machine, but the wear occurring *during* the first thousand, second thousand, and third thousand revolutions, that is to say to consider the first differences of the original data.

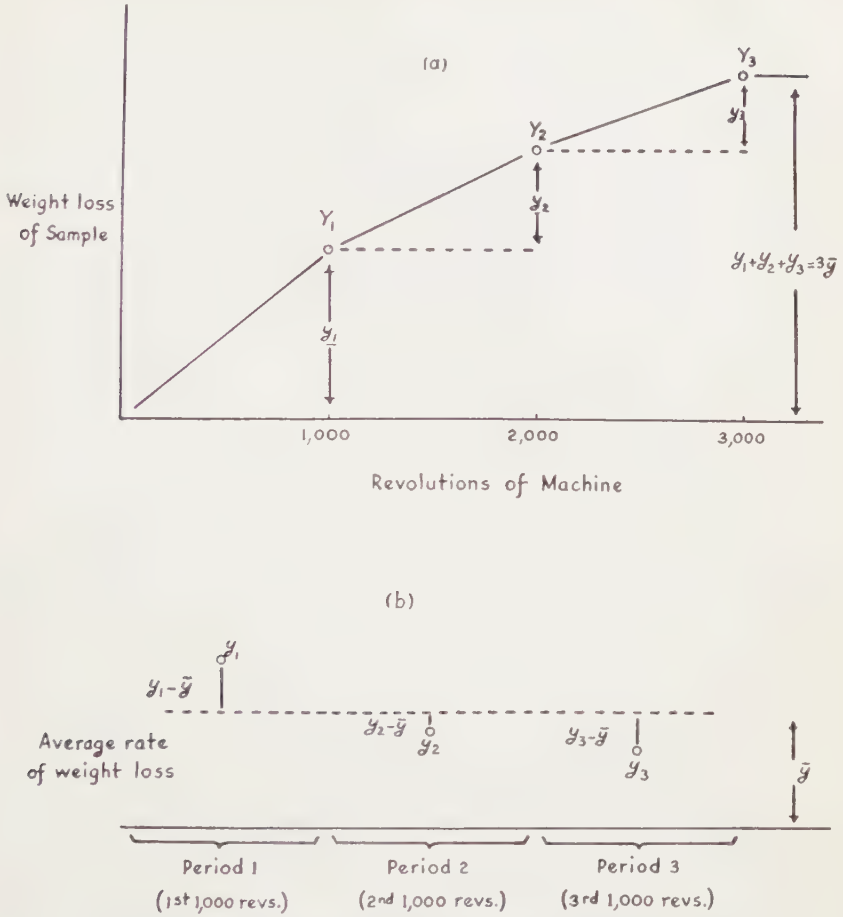


FIGURE III. ANALYSING FIRST DIFFERENCES OF WEAR DATA

In Fig. IIIa, Y_1, Y_2 and Y_3 are the total weight losses for a particular specimen at 1000, 2000, and 3000 revolutions of the machine. We consider $y_1 = Y_1$, $y_2 = Y_2 - Y_1$ and $y_3 = Y_3 - Y_2$; y_1, y_2 and y_3 can be regarded as measuring the average *rates* of wear in milligrams per 1000 revolutions during the three periods, and $\bar{y}$, the mean as measuring the *overall* average rate. Since $3\bar{y}$ is equal to Y_3, $\bar{y}$ is proportional to the total wear during the experiment. The three periods of wear considered can then formally be regarded as a further factor, "periods", and the curve obtained by plotting the differences y_1, y_2, y_3 (Fig. IIIb) indicates the approximate shape of the wear *rate* curve. Now whatever other information is required from the data it will usually be the case that the overall effects of

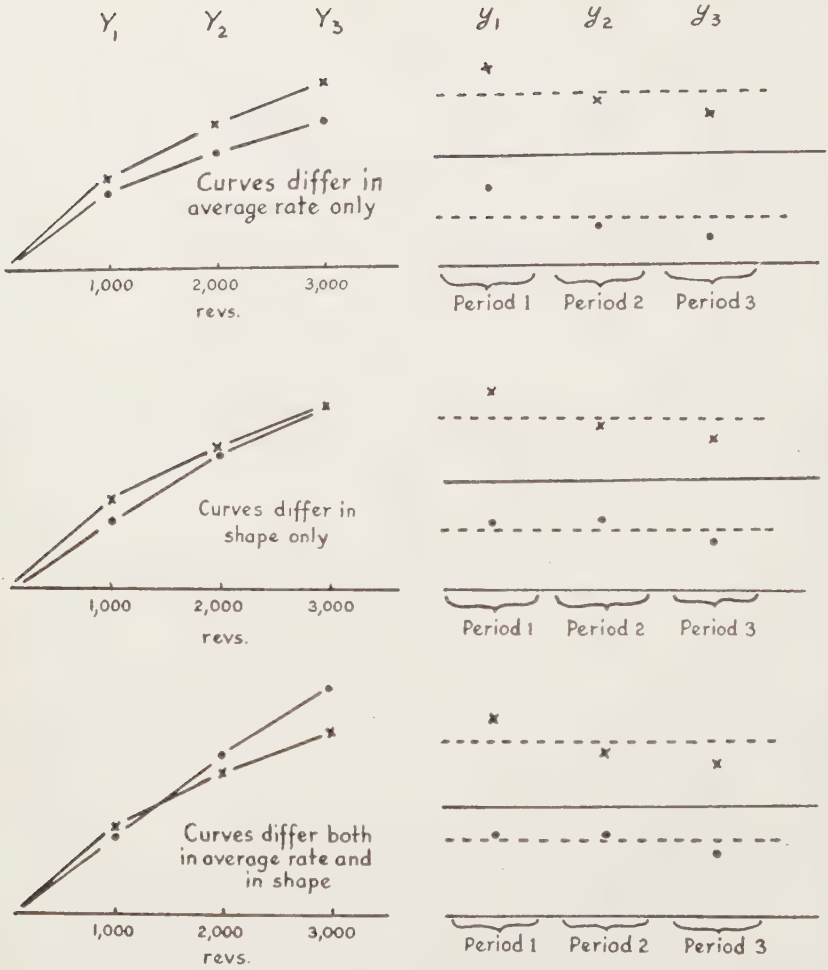


FIGURE IV. COMPARISON OF CURVES USING AVERAGE RATE AND "SHAPE"

treatments will be of major interest and these effects can be elucidated by analyzing the variate $\bar{y}$; it may be, however, that the treatments have effected not only the average rates, but have also influenced the configuration of the individual period rates about their averages, and such effects as this can be regarded as interactions between the factor concerned and "periods".

If (Fig. IV) a factor affects only the average wear rate $\bar{y}$, i.e. alters y_1 , y_2 , and y_3 , equally, (and there is consequently no interaction with "periods"), we shall say that the wear rate curve has been altered only

in level, but if the deviations $y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}$ are affected, we shall say that the wear rate curve is also altered in *shape*. This use of the average level $\bar{y}$ and the deviations $y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}$, to describe the curve is analogous to the size and shape analysis used in discrimination problems by Penrose (1947) and Smith (1947). On this definition, then, two curves are said to have different shapes when the wear rates differ by different amounts at different stages of wear.

The data for Fig. I, after differences have been taken, are set out in Table A. The figures in the table corresponding with periods 1, 2, and 3 refer to the wear between 0–1000, 1000–2000, 2000–3000 revolutions of the machine respectively.

TABLE A

WEAR OF COATED FABRICS IN MILLIGRAMS

Surface Treatment	Filler	Proportion of Filler								
		Q_1 (25%)			Q_2 (50%)			Q_3 (75%)		
		Period			Period			Period		
		1	2	3	1	2	3	1	2	3
T_0	F_1	194	192	141	233	217	171	265	252	207
		208	188	165	241	222	201	269	283	191
	F_2	239	127	90	224	123	79	243	117	100
		187	105	85	213	123	110	226	125	75
T_1	F_1	155	169	151	198	187	176	235	225	166
		173	152	141	177	196	167	229	270	183
	F_2	137	82	77	129	94	78	155	76	91
		160	82	83	98	89	48	132	105	67

2.1 A mathematical model for the experiment.

The error variances for the data are as follows:

Revolutions	Error Variance	Revolutions	Error Variance
0–1000	269	0–1000	269
0–2000	456	1000–2000	200
0–3000	935	2000–3000	222

A characteristic of the cumulative data is seen to be the increase of the error variance as the test proceeds, the variance of the differences however, remains stable from period to period. This is to be expected, for much of the variation probably arises in the operation of the testing machine itself and the variation occurring in equal periods of running might be expected to be the same. Also, in the original data the errors would be expected to be correlated from one observation to the next, since for example abnormally low wear occurring in the first 1000 revolutions would be reflected in subsequent values. So far as the variation of the machine was concerned however, correlation between wear in successive *periods* might be expected to be much smaller and possibly negligible.

For the moment then, we will assume that the departure of y_{ti} (the loss in weight during the i^{th} period of wear of a sample having the t^{th} factor combination) from its mean value η_{ti} can be represented by two independent random variables ϵ_t and δ_i each being normally distributed about zero, the first with variance σ_1^2 allows for *overall* variation in mean rate in duplicate specimens and the second with variance σ_0^2 allows for variations associated with individual periods.

$$y_{ti} - \eta_{ti} = \epsilon_t + \delta_i \tag{1}$$

If this is true, then the analysis of variance for the data in Table A, will be analogous to that for a split-plot agricultural experiment, the "plots" being the samples of material and the three values for wear in successive periods corresponding to the splitting of the plots. As with split plot experiments, the analysis will consist of two parts, each part having its own error estimate. This analysis for the wear data is set out in Table B. The entries in the table are the mean squares corresponding to the effects; the figures in brackets refer to the degrees of freedom available for the comparisons.

The mean squares in the left-hand column are obtained after averaging over "periods" and they enable hypotheses concerning the average wear rate (or what is equivalent, the overall wear) to be tested; the correction for the mean, denoted by I, is included for completeness. Significance is judged by comparison with the mean square of 312 at the foot of this column; this error term has 12 degrees of freedom and is an estimate of the "between samples" variance $3\sigma_1^2 + \sigma_0^2$ as is each of the mean squares in this column if the treatments are without effect. The asterisks indicate significance at the 5, 1 and 0.1% points respectively.

The mean squares in the right-hand column correspond to interactions with "periods" and significant interactions imply that a factor has

TABLE B
ANALYSIS OF VARIANCE FOR THE DATA OF TABLE A

Source		Averaged over "periods" (level of rate curve)		Interactions with "periods" (Shape of rate curve)	
Mean	(I)	(1)	1,866,956	(2)	30,479***
Main effects					
% Filler	(Q)	(2)	6,785***	(4)	440
Type of Filler	(F)	(1)	107,803***	(2)	9,144***
Surface Treatment	(T)	(1)	24,494***	(2)	4,124***
Interactions					
$Q \times F$		(2)	4,942***	(4)	354
$Q \times T$		(2)	397	(4)	172
$F \times T$		(1)	1,682*	(2)	1,164***
$Q \times F \times T$		(2)	150	(4)	116
Error		(12)	312	(24)	190

affected the quantities $y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}$, that is, that it has altered the "shape" of the rate curve. The interaction of I with the periods factor P is $P \times I = P$ the main effect for periods, which indicates whether the mean *rate* of growth has remained constant from period to period i.e. whether the mean growth curve is a straight line. The error mean square of 190 appropriate for testing these effects has 24 degrees of freedom and is an estimate of σ_0^2 the "within samples" variance, as is each of the mean squares in this column on the hypothesis that the treatments are without effect. The entries in this column can be most easily calculated by first carrying through an analysis of variance for each period separately; and then using the identity;

$$\left\{\begin{matrix}\text{Sum of squares}\\ \text{for interaction}\\ \text{with periods}\end{matrix}\right\} = \left\{\begin{matrix}\text{total of sums}\\ \text{of squares for}\\ \text{individual periods}\end{matrix}\right\} - \left\{\begin{matrix}\text{sums of squares}\\ \text{for average}\\ \text{effects.}\end{matrix}\right\}$$

For example, for the surface treatment T the sums of squares for the three periods of wear were found to be 26,268, 5,017 and 1,457 respectively; the sum of squares for the average effect was 24,494, and the sum of squares for interaction with periods was therefore $26{,}268 + 5{,}017 + 1{,}457 - 24{,}494 = 8{,}248$ and the mean square, 4,124 as shown in the table; all the entries in this column including, of course, the error term, are calculated in this way. This procedure can be followed whether the

remaining effects are orthogonal or not. For example, in the balanced incomplete block design used for the assessment of tire treads, the usual (nonorthogonal) analysis for incomplete block designs is first carried through for the means, averaging over periods; this analysis is then repeated for each period separately and the above identity used to determine the interactions with periods.

2.2 Interpretation of the analysis.

Considering first the left-hand column of table B, we see that all the main effects are highly significant and that interactions exist between Q and F, that is, between the proportion of filler and the type of filler used and between F and T, the type of filler and the surface treatment. The appropriate tables of mean values showing the nature of these interactions follow:

		% FILLER (Q)		
		25	50	75
Type of Filler	F_1	169	199	231
	F_2	121	120	126

		SURFACE TREATMENT	
		T_0	T_1
Type of Filler	F_1	214	145
	F_2	186	99

From the first table, we see that the average rate of wear is less with the second filler, and this average rate was virtually unaffected by the increase in the percentage of filler, whereas with the first filler, the material wears more, and the wear increases as the percentage is increased. From the second table of means we see that the beneficial effect of the surface treatment is even more pronounced with the second than with the first filler.

So far, we have considered the effect of factors only on the average rate of wear, we now need to consider whether these effects change at different stages of wear; that is to say whether the factors affect the shape of the rate curves. To do this, we consider the mean square for the interaction of each of the effects with periods, shown in the right-hand

column of the table; the significance of these items is judged by comparison with the error term at the foot of this column. The % filler (Q) and % filler × type of filler (QF) effects have no significant interactions with periods, so we can regard our conclusions for these effects as probably true at all stages of wear; however both F (the type of filler) and T (the surface treatment) show strong interactions. The tables of mean values are as follows:

		Period 1	Period 2	Period 3
Type of Filler	F_1	215	213	172
	F_2	181	104	82
		Period 1	Period 2	Period 3
Surface Treatment	T_0	231	173	134
	T_1	165	144	119

From the first table we conclude that not only do the samples having filler (1) wear at a greater rate than those having filler (2), but also the fall off in wear is greater with filler (2). From the second table we conclude that although the surface treatment has a favorable effect at all stages of wear this is (as would be expected) most marked initially. Finally, we see that the interaction $F \times T$, between type of filler and surface treatment, interacts with periods and, on consulting a table of mean values, this is seen to be due to the fact that the interaction between filler and surface treatment found before, is confined to the first period of wear alone.

The conclusions from the experiment can therefore be set out as follows:

1. Filler (2) results in less wear than filler (1), the rate of fall off of wear is greater and the protective effect of surface treatment is more marked with (2) than with (1).

2. Whereas increasing the % of filler (1) results in increased rate of wear, the % of filler (2) can be increased, at least to 75%, without increasing wear.

3. The surface treatment markedly reduces wear especially during the early stages and when filler (2) is used.

It is clear that the two fillers are behaving differently and in a full analysis the data would be split into two, and an analysis made for each filler separately. We shall not however elaborate the analysis further

here, since our purpose is merely to show that when the set-up given in equation (1) can be regarded as valid, an exceedingly simple and informative analysis can be made. We now proceed to show how it is possible to test whether significant departure from the simple set-up occurs or not.

3. A TEST FOR DEPARTURE FROM THE SIMPLE MODEL

Instead of equation (1) let us write

$$y_{ti} - \eta_{ti} = z_{ti} \tag{2}$$

Then if the simple set-up is valid

$$z_{ti} = \epsilon_t + \delta_i \tag{3}$$

and z_{ti} would be distributed in a three-variate multinormal distribution with each variance equal to $\sigma_1^2 + \sigma_0^2$ and each covariance equal to σ_1^2. (For simplicity the test is illustrated for three variates; it can of course be immediately generalized to the p-variate case.) To check the validity of our analysis therefore we must test the hypothesis, that all the variances are equal and all the covariances are equal; that is, that V*, the matrix of variances and covariances is of the form:

$$V_0 = \begin{bmatrix} a & d & d \\ d & a & d \\ d & d & a \end{bmatrix} \tag{4}$$

against the alternative that the variances are not all the same and the covariances not all the same; that is, that the variance covariance matrix is of the form:

$$V_1 = \begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} \tag{5}$$

where a, b, c are not all equal and d, e, f, are not all equal.

The null hypothesis of (4) is a little more general than that implied by equation (3) for negative values of the covariance d are possible with (4) (although since V_0 must be positive definite d cannot be less than $-a/(p-1)$), whereas correlation arising from a common component ϵ_t in (1) must clearly be positive. However, an analysis of the type given in Table B is valid for any positive definite matrix of form (4) so that this extension is appropriate.

3.1 The Test Criterion

A criterion for testing a statistical hypothesis of this form has been

*We tacitly assume here that the variance covariance matrix V remains constant from group to group, that is to say is itself unaffected by the treatments; we consider this assumption later in § 7.

obtained by Wilks (1946) using the likelihood ratio method of Neyman and Pearson (1928). Let $c_{ij} = c_{ji}$ denote the error sums of squares and products calculated from the sample for the i^{th} and j^{th} variates, then the criterion is

$$\Lambda = \frac{\Delta_1}{\Delta_0} \frac{\begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix}}{\begin{vmatrix} \bar{c}_{ii} & \bar{c}_{ij} & \bar{c}_{ij} \\ \bar{c}_{ij} & \bar{c}_{ii} & \bar{c}_{ij} \\ \bar{c}_{ij} & \bar{c}_{ij} & \bar{c}_{ii} \end{vmatrix}} \tag{6}$$

where $\bar{c}_{ii}$ is the average sum of squares $(c_{11} + c_{22} + c_{33})/3$ or in general $(\sum_i^p c_{ii})/p$ and $\bar{c}_{ij}$ is the average sum of products $(c_{12} + c_{13} + c_{23})/3$, or in general

$$\left\{ \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} c_{ij} \right\} \Big/ \left\{ \frac{1}{2} p(p-1) \right\}$$

and p is the number of variates (3 in this case). The value of the determinant Δ_0 in the denominator can be easily shown to be equal to $[\bar{c}_{ii} + (p-1)\bar{c}_{ij}][\bar{c}_{ii} - \bar{c}_{ij}]^{p-1}$ which simplifies the calculations.

For the wear test example the sums of squares and products for error can be most easily calculated from the differences between duplicate observations in table A. These differences are set out in table C, for example, $14 = 208 - 194$, $-4 = 188 - 192$, etc.

TABLE C

DIFFERENCES BETWEEN DUPLICATES

Period	T_0						T_1					
	F_1			F_2			F_1			F_2		
	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
1	14	8	4	−52	19	−17	18	−21	−6	23	−31	−23
2	−4	5	31	−22	0	8	−17	9	45	0	−5	29
3	24	30	−16	−5	31	−25	−10	−9	17	6	−30	−24
Total	34	43	19	−79	50	−34	−9	−21	56	29	−66	−18

The sums of squares and products are then obtained as follows:

		Period			
		1	2	3	S
Period	1	3225.0	−80.5	1656.5	4801.0
	2	−80.5	2405.5	−112.0	2213.0
	3	1656.5	−112.0	2662.5	4207.0

For example $3225.0 = [14^2 + 8^2 + 4^2 + \cdots + (-23)^2]/2$

$$-80.5 = [14 \times (-4) + 8 \times 5 + \cdots + (-23) \times 29]/2.$$

The divisor 2 is employed since the values concerned are differences between two observations. The check column S shows the sum of products between each of the variates and the column totals of table C; this supplies an independent check on the working, for example, for the second row of the matrix

$$-80.5 + 2405.5 - 112.0 = 2213.0.$$

The value of the determinant of this matrix which is the numerator Δ_1 of the criterion is found to be $14{,}026 \times 10^6$. To calculate the denominator Δ_0 of the criterion, we find:

$$\bar{c}_{ii} = 2764.3, \qquad \bar{c}_{ij} = 488.0,$$

$$\bar{c}_{ii} + 2\bar{c}_{ij} = 3740.3 \qquad \bar{c}_{ii} - \bar{c}_{ij} = 2276.3$$

and $\Delta_0 = 3740.3 \times (2276.3)^2 = 19{,}381 \times 10^6$, whence $\Lambda = 0.7237$

3.2 The test of significance

We now wish to test whether or not this value of Λ is exceptionally small. The exact distribution is not known in the general case; however, an expression for the moments of Λ has been given by Wilks (1946) and for these, sufficiently accurate approximations can be calculated. The present author (Box 1949) has given a general distribution theory for a very wide class of what may be called "Λ" statistics, whose moments can be written in the form

$$E(\Lambda^a)^h = \text{constant} \cdot \left[\frac{\prod_{j=1}^{k} (y_j^{y_j})}{\prod_{i=1}^{m} (x_i^{x_i})} \right]^h \frac{\prod_{i=1}^{m} [\Gamma\{x_i(1+h) + \xi_i\}]}{\prod_{j=1}^{k} [\Gamma\{y_j(1+h) + \eta_j\}]} \tag{7}$$

Given the value of k, m, and the x_i, ξ_i, y_i and η_i, general formulae are provided from which taking $M = 2a \log_e \Lambda^{-1}$ as a working statistic, an accurate χ^2 series solution can be obtained, and simple χ^2 and F approximations. For the statistic here considered, Wilks' expression for the moments can be written

$$E(\Lambda^{\nu/2})^h = \text{constant} \cdot (p-1)^{\frac{1}{2}(p-1)\nu h} \frac{\prod_{i=1}^{p-1}\left[\Gamma\left\{\frac{\nu}{2}(1+h) - \frac{i}{2}\right\}\right]}{\Gamma\left\{\frac{\nu}{2}(p-1)(1+h)\right\}} \tag{8}$$

ν being the degrees of freedom of the sums of squares and products tested. This is seen to be of the same form as (7) so the general theory can be applied.

Making the substitutions we find we should take for our working statistic

$$M = \nu \log_e \Lambda^{-1}$$

The more the data depart from the simple set-up, the smaller will be Λ and the greater the value of M. To test whether M is large enough to indicate a "significant" departure from the simple set-up we calculate

$$f_1 = (p^2 + p - 4)/2, \qquad A_1 = \frac{\{p(p+1)^2(2p-3)\}}{\{6\nu(p-1)(p^2+p-4)\}}$$

and refer $(1 - A_1)\, M$ to tables of χ^2 with f_1 degrees of freedom. An approximation which is rather more precise especially when p is large and/or ν is small is supplied by calculating

$$A_2 = \frac{(p-1)p(p+1)(p+2)}{6\nu^2(p^2+p-4)}, \quad f_2 = \frac{f_1+2}{A_2 - A_1^2}, \quad b = \frac{f_1}{1 - A_1 - f_1/f_2}$$

and referring M/b to tables of the F distribution with f_1 and f_2 degrees of freedom.

In the present example $\nu = 12$ and $p = 3$. Whence $f_1 = 4$, $A_1 = 0.125$, $A_2 = 0.0174$, $M = 12 \log_e (1/0.7237) = 3.880$. $(1 - A_1)\, M = 3.395$ and this quantity is to be referred to tables of χ^2 with four degrees of freedom, from which we can conclude at once that on this data there is no reason to question the simple set-up. The actual probability given by this approximation for the chance occurrence of a value of M as great or greater than this is slightly less than 0.5. Using the F approximation a very similar result would have been obtained, we find

$$f_1 = 4, \qquad f_2 = 3456, \qquad b = 4.578.$$

Thus $M/b = 0.8476$ is to be referred to tables of F with 4 and 3456 degrees of freedom, and consulting the table of Thompson and Merrington (1943) we again find a value for the probability of chance occurrence slightly below 0.5. In this particular case both approximations are quite accurate, and the values they give are almost identical. When p is larger and ν is smaller the χ^2 approximation is less accurate and the F approximation differs more markedly from it, f_2 being no longer very large as it is in this example.

4. AN ALTERNATIVE SET-UP

We have seen how, by the device of taking differences, the problem of interpreting wear data was facilitated and a simple analysis was possible if it was reasonable to assume a particular set-up. We have shown that for a particular example the hypothesis that the set-up was of this simple form was not contradicted by the data.

The simple set-up would be expected to represent wear data satisfactorily if most of the variation arose from the operation of the machine itself or if the variation within replicate specimens of material were mainly confined to changes in *average* abrasion resistance and not to changes in "shape" of the wear curves which might give rise to serial correlation between differences. Observational errors would also tend to cause departures from the set-up, for an apparent increase in wear rate during one period would be compensated by an apparent decrease in the next, and this would lead to negative correlation between successive differences. In some cases therefore we would not expect the variances and covariances, even after differencing of the data, to be capable of adequate representation by the simple pattern of equation (4) and the more general set-up typified by equation (5) would have to be adopted.

When a research program is being carried out in which the results will appear in the form of wear curves or growth curves, it is worthwhile paying particular attention, in the preliminary experiments, to the form of the variance co-variance matrices, so that decisions may be reached concerning a set-up, and method of analysis, suitable for use in this *particular* investigation; that is to say with this particular type of material and interval between observations. Consider now the rat growth data shown in detail in Table D at end of article, which was plotted in Figure II. For the reasons given above we should not expect the weight gains for these animals, even after differencing, to be uncorrelated from one period to another. In fact, if we denote the five variates recording initial weight and weight after one, two, three, and four weeks by Y_0, Y_1, Y_2, Y_3, Y_4 and the differences, $Y_1 - Y_0$, $Y_2 - Y_1$, $Y_3 - Y_2$, $Y_4 - Y_3$ by y_1, y_2, y_3, y_4, the matrix of sums of squares and products for the 24 error degrees of freedom for y_1, y_2, y_3 and y_4, is found to be

$$\begin{bmatrix} 582.3 & 42.5 & -55.5 & -74.6 \\ 42.5 & 609.0 & 626.5 & 344.5 \\ -55.5 & 626.5 & 1046.7 & 459.0 \\ -74.6 & 344.5 & 459.0 & 853.0 \end{bmatrix} \tag{9}$$

From which proceeding as before we find

$$\Lambda = 0.3534, \qquad M = 25.0, \qquad 1 - A_1 = 0.9277, \qquad f = 8.$$

Referring $(1 - A_1)M = 23.2$ to tables of χ^2 with 8 degrees of freedom we conclude that the probability of a chance deviation from the simple set-up as great as this is less than 0.01; a similar result is given by the F approximation and we are therefore led to reject this set-up.

Although differencing is not completely successful in transforming the data into a form in which the variances are equal and the covariances are equal, it is worth noting that the differences do show a great deal more uniformity in this respect than the data before differencing. The corresponding error matrix for sums of squares and products for the overall weight gains $Y_1 - Y_0$, $Y_2 - Y_0$, $Y_3 - Y_0$, and $Y_4 - Y_0$ is

$$\begin{bmatrix} 582.3 & 624.8 & 569.3 & 494.7 \\ 624.8 & 1276.3 & 1847.3 & 2117.2 \\ 569.3 & 1847.3 & 3465.0 & 4193.9 \\ 494.7 & 2117.2 & 4193.9 & 5775.8 \end{bmatrix} \tag{10}$$

and the value of χ^2 obtained on applying the test to these values is 109.1; the transformation has thus gone a good deal of the way towards bringing the data to the simpler form.

In making tests of an actual set-up, it should be borne in mind that the important consideration is how far departures from the assumptions made will affect the *tests based on these assumptions*. This problem has been investigated in a number of cases by the present author and it is hoped to publish the results elsewhere in the near future. It appears that minor departures of the data from independence and homoscedasticity of the type considered here will not seriously influence subsequent tests of significance; a similar conclusion was reached by Daniels (1938). Thus, in the wear curve example we found no significant departure from the simple set-up, and although this did not mean that no such depar-

ture could have occurred, but possibly, merely that the data were not sufficiently extensive for it to be detected, it is probable that little error was made by adopting the simple set-up in this case. In the growth curve example however a marked departure is apparent and it would be safer to adopt the less restrictive assumptions of the alternative hypothesis of equation (5).

We shall assume that the variates Y_0, Y_1, Y_2, Y_3 and Y_4 are distributed about their mean values in a 5-variate normal distribution, the same for each rat, and this of course implies that any set of variates derived from these, by linear transformation, will also be distributed multinormally; in particular the differences will be distributed in that form.

4.1 The Multivariate Test

Now if a single variate only, say the overall increase in weight, were being analyzed, the hypothesis concerning the significance of treatment differences could be tested by means of the analysis of variance, that is to say the criterion

$$\frac{\text{mean square for treatments}}{\text{mean square for error}}$$

would be referred to tables of the F distribution with the appropriate numbers of degrees of freedom. Alternatively (see for example Kolodziejczyk 1935) the criterion*

$$\Lambda = \frac{\text{sum of squares for error}}{\text{sum of squares for error} + \text{sum of squares for treatment}}$$

could be employed, and the test carried out by referring to tables of the incomplete B-function (Karl Pearson; 1934, Thomson; 1941), the result of course would be precisely the same.

For our present purpose the latter criterion is of more interest, because it can be directly generalized (Wilks; 1932, Pearson and Wilks; 1933, Bartlett; 1934, 1938) to the case where the observations are not single variates but multivariates, whereas the former criterion cannot. Thus the hypothesis that the mean value for each of the variates is the same from group to group; (for example, that the gains in weight during the first week are all equal, *and* the gains during the second week are all equal, etc.) can be tested by calculating the criterion;

*Λ is used in the paper to denote a criterion of the form associated with the likelihood ratio method of Neyman and Pearson. M is used to denote a logarithmic statistic derived from Λ. The likelihood statistic Λ and the derived quantity M referred to in this section are of course different from the criterion discussed in § 3.

$$\Lambda = \left| \frac{\text{sums of squares and products for error}}{\text{sums of squares and products for error} + \text{sums of squares and products for treatment}} \right|$$

where *determinants* whose elements are sums of squares and products replace the single sums of squares of the univariate criterion.

Thus, had we desired to use the more general set-up in the wear test example, an *overall* test for each of the main effects and interactions could have been applied by calculating the 3×3 matrix of sums of squares and products first for error and then for the particular main effect or interaction concerned, and hence calculating the criterion Λ. This test would not have distinguished between changes in average and changes in shape but would have been an overall test including both.

The exact distribution of Λ is known only for certain special cases, however, this is another of the general class of statistics whose moments can be written in the form of equation (7), and simple approximations which are perfectly general and which are usually sufficient for most practical purposes can be obtained. To preserve generality, even when the exact distribution is available, these approximations will be used in all the tests that follow. If n is the number of degrees of freedom for *treatments plus error*, q the number of degrees of freedom for treatments and p the number of variates then Bartlett's (1938) χ^2 approximation is obtained by calculating

$$M = n \log_e \Lambda^{-1} \qquad A_1 = (p + q + 1)/2n \qquad f_1 = pq$$

and referring $(1 - A_1)M$ to tables of χ^2 with f_1 degrees of freedom. This approximation tends to break down if n is small or p and q are large and in these cases it is worthwhile calculating the more accurate F approximation (Box 1949). For this we calculate in addition

$$f_2 = \frac{12n^2(pq + 2)}{p^2 + q^2 - 5}, \qquad b = \frac{pq}{1 - A_1 - f_1/f_2},$$

and refer M/b to tables of F with f_1 and f_2 degrees of freedom.

In the growth curve example if we consider the variates y_1, y_2, y_3, y_4 (that is, the first differences of the weight gains) the matrix for sums of squares and products for treatments is found to be

$$\begin{bmatrix} 81.7 & 37.2 & 11.5 & 112.9 \\ 37.2 & 476.9 & 782.7 & 787.4 \\ 11.5 & 782.7 & 1315.9 & 1260.1 \\ 112.9 & 787.4 & 1260.1 & 1334.0 \end{bmatrix} \tag{11}$$

The corresponding matrix for sums of squares and products for error has already been given (9). The error plus treatment matrix is obtained by adding each element in the error matrix to the corresponding element in the treatment matrix. The ratio Λ of the error determinant to the error plus treatment determinant is then found to be 0.2661, and $n = 26$, $p = 4, q = 2$. For such large values of n and comparatively small values of p and q Bartlett's χ^2 approximation will be adequate and we find

$$M = 34.4, \qquad A_1 = 7/52, \qquad f_1 = 8$$

and referring $(1 - A_1)M = 29.8$ to tables of χ^2 with 8 degrees of freedom we conclude that the mean values for y_1 , y_2 , y_3 , and y_4 representing the growth during the first, second, third, and fourth weeks differ very significantly from group to group ($P < 0.001$).

4.2 Special Properties of the Criterion.

Before proceeding further we note certain important properties of this criterion used in the multivariate extension of the analysis of variance.

1. The criterion is invariant under non-singular linear transformation of the variates. Thus, in the example above, if we had analyzed the total gains in weight instead of the differences, or had applied any other linear transformation of this sort to the data, the value of the overall criterion would have been unchanged.

2. The sums of squares and products matrix for any *new* set of variates obtained by linear transformation can be found directly by applying the transformation to the rows and columns of the matrix of sums of squares and products of the *old* set of variates. For example if the matrix (10) for total weight gains were known, (9) the corresponding matrix after differencing the data, could be obtained by applying the differencing process to the rows and columns of (10) itself; it is not necessary to make the transformation to the original data and recalculate.

3. In the calculation of determinants, the method of pivotal condensation (see for example Aitken, 1948) provides a rapid practical procedure; *this device also provides a useful method for the elimination of variables.* As an example consider a determinant Δ of sums of squares and products for, say, three variates y_1 , y_2 , y_3 ,

$$\Delta = \begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} \qquad \text{where } c_{ij} = c_{ji}$$

dividing through the first row by c_{11} we obtain a new first row

$$1 \quad \frac{c_{12}}{c_{11}} \quad \frac{c_{13}}{c_{11}}$$

if this is subtracted c_{12} times from the second row and c_{13} times from the third, we have

$$\Delta = c_{11} \begin{vmatrix} 1 & \frac{c_{12}}{c_{11}} & \frac{c_{13}}{c_{11}} \\ 0 & c_{22} - \frac{c_{12}^2}{c_{11}} & c_{23} - \frac{c_{11}c_{13}}{c_{11}} \\ 0 & c_{23} - \frac{c_{12}c_{13}}{c_{11}} & c_{33} - \frac{c_{13}^2}{c_{11}} \end{vmatrix}$$

and writing $c_{22} - c_{12}^2/c_{11}$ as $c_{22.1}$, $c_{23} - c_{12}c_{13}/c_{11}$, as $c_{23.1}$, etc. and expanding the determinant along the first column, we have

$$\Delta = c_{11} \begin{vmatrix} c_{22.1} & c_{23.1} \\ c_{23.1} & c_{33.1} \end{vmatrix}$$

that is

$$\Delta = \Delta_{123} = c_{11}\, \Delta_{23.1} .$$

As is well known, to compute the value of any $p \times p$ determinant, the process may be repeated $p - 1$ times till the determinant is reduced to the product of p known quantities, and this process is the basis of the Gauss-Doolittle method for the solution of the linear equations, and can be still further simplified (see for example Dwyer; 1942). What is interesting for our purpose is the fact that the elements of $\Delta_{23.1}$ are the sums of squares and products for y_2 and y_3 after eliminating the variable y_1, that is to say they are the sums of squares and products of deviations from the regressions of y_2 and y_3 on y_1. Now if condensation of this sort is applied simultaneously to numerator and denominator of Λ we have

$$\Lambda = \frac{c_{11}\ (\text{error})}{c_{11}\ (\text{error} + \text{treatments})} \quad \frac{\Delta_{23.1}\ (\text{error})}{\Delta_{23.1}\ (\text{error} + \text{treatments})}$$

$$= \Lambda_1\, \Lambda_{23.1}$$

In the above equation Λ_1 corresponds to a univariate analysis of variance for the variable y_1 and the second component to a multivariate analysis of covariance for the remaining variables with the first variate y_1 eliminated. Any number of variables can be eliminated in this way,

the number of degrees of freedom for error being correspondingly reduced after each elimination. A successive elimination of variates of this sort was applied by Bartlett to the linear quadratic and cubic components fitted to the growth curves of pigs by Wishart (1939).

4.3 Further Analysis of the Data.

Now even though the taking of differences does not result in a simplification of the set-up, it allows the changes in the wear curves to be more easily appreciated and may still be employed in the interpretation of the overall criterion. We shall therefore again consider the mean growth rate $\bar{y}$ and the deviations from the mean $y_1 - \bar{y}$, etc. Only three of the four deviations from the mean are linearly independent and all the information concerning departures from the mean is contained in any three of them, we therefore consider the variates $\bar{y}$, $y_1 - \bar{y}$, $y_2 - \bar{y}$ and $y_3 - \bar{y}$; $y_4 - \bar{y}$ is omitted from the analysis, (exactly the same result will be obtained whichever of the deviations is omitted). Using the second property noted above, the entries for sums of squares and products for the new variates are obtained by direct transformation. For example we obtain the error matrix for the new variates from the error matrix (9) for y_1 , y_2 , y_3 , and y_4 as follows; first applying the transformation to the rows, corresponding to the operation of taking the mean $\bar{y}$ we replace the elements of the first row of (9) by their column means, the second row is then obtained by subtracting these values from the elements of the first row of (9) corresponding to the operation of taking $y_1 - \bar{y}$; the third and fourth rows are found similarly, and the whole set of operations is then carried out on the columns. A partial check is supplied by the symmetry of the final transformed matrix and a complete check can be made by calculating the sums for each row and column of the final matrix and confirming that these totals agree with the values found by operating on the sums of rows and columns of the original matrix. From the error matrix (9) we obtain

	$\bar{y}$	$y_1 - \bar{y}$	$y_2 - \bar{y}$	$y_3 - \bar{y}$
$\bar{y}$	361.0	−237.3	44.6	158.2
$y_1 - \bar{y}$	−237.3	695.9	−125.8	−337.4
$y_2 - \bar{y}$	44.6	−125.8	158.7	62.7
$y_3 - \bar{y}$	158.2	−337.4	62.7	369.3

and applying the same procedure to the error plus treatment matrix we have

	$\bar{y}$	$y_1 - \bar{y}$	$y_2 - \bar{y}$	$y_3 - \bar{y}$
$\bar{y}$	935.5	−751.0	−8.8	426.2
$y_1 - \bar{y}$	−751.0	1230.5	−96.0	−654.7
$y_2 - \bar{y}$	−5.8	−96.0	168.0	56.3
$y_3 - \bar{y}$	426.2	−654.7	56.3	574.6

The Λ criterion for means alone is therefore:

$$\Delta(\bar{y}) = \frac{360.9}{935.4} = 0.3859$$

and for reasons already given we shall employ Bartlett's approximation to make the test of significance. We find $(1 - A_1)M = 22.9$, should be referred to χ^2 tables with two degrees of freedom whence we deduce that the mean growth rates differ very significantly ($P < .001$) from group to group. To test the deviations from the means, that is to test whether the "shape" of the growth curve varies from group to group we calculate the ratio of the 3 x 3 determinant for error to that for error + treatments for the three variates $y_1 - \bar{y}$, $y_2 - \bar{y}$, $y_3 - \bar{y}$. We find

$$\Lambda(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}) = 0.4366$$

and $(1 - A_1)M = 19.0$ is referred to tables of χ^2 with 6 degrees of freedom. This value is significant ($P < .01$) and we therefore conclude that, not only the mean level, but also the shape of the curve is changing from group to group. A table of mean values indicates the nature of the differences.

MEAN GAINS IN WEIGHT (GRAMS)

Period	Group 1	Group 2	Group 3
1st week	24.5	20.3	21.6
2nd week	27.5	29.0	19.5
3rd week	24.1	29.3	12.4
4th week	30.5	30.1	15.8
Mean rate (grams/week)	26.7	27.2	17.3

Further tests show that no significant differences occur between groups 1 and 2, i.e., that the treatment of group 2 is without significant

effect, however group 3 differs from the other groups both in average level and in shape. In the first two groups we find a fairly steady rate, which if anything, is tending to increase, whereas a fall in growth rate is found in the third group.

So far the average effects and "shape" effects have been treated separately, but it may be relevant to inquire whether the effects found in the two parts of the analysis can be regarded as separate entities, or whether they are really manifestations of the same thing. In Fig. II it is noticeable that the growth curves of group 3 not only show a low average rate, but also tend to be convex upwards whereas the growth curves of groups 1 and 2 have a higher average and are if anything concave. Now there may also be a tendency *within* the groups, for these curves with low average rates of growth to be also those which are most convex; we may therefore wish to test whether *given the change in mean growth rate*, any differences in "shape" occur, other than would be expected from the internal evidence of the groups concerning the relation between "shape" and mean value. To make the test we use the third property of the Λ criterion mentioned above; the criterion for variables $y_1 - \bar{y}$, $y_2 - \bar{y}$, $y_3 - \bar{y}$ *given* $\bar{y}$ is calculated by dividing the overall criterion for the 4 variables by that for the single variable $\bar{y}$.

$$\Lambda(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y} : \bar{y}) = \Lambda(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}, \bar{y})/\Lambda(\bar{y})$$

$$= \Lambda(y_1 , y_2 , y_3 , y_4)/\Lambda(\bar{y})$$

$$= \frac{0.2661}{0.3859} = 0.6896$$

Since one variable $\bar{y}$ has been eliminated we have $n = 25$, $q = 2$, $p = 3$, and $(1 - A_1)M = 8.18$ is referred to tables of χ^2 with 6 degrees of freedom. The probability of chance occurrence of such a value is about 0.25. We see, therefore, that there is no evidence of differences in shape other than would be expected from the *internal* relation between average and shape.

5. REDUCTION OF THE DATA

In some experiments, weighings are made at very short intervals of time, and the number of points p' for each growth curve is large. Usually however the salient features of the curves will be described by employing fewer than p' constants. Thus to reduce his data Wishart (1938) fitted orthogonal polynomials up to the third degree to the overall weight gain for each of a number of pigs receiving different rations, and analyzed the regression constants. The essential idea is to reduce the

data without sacrificing the extra precision given by the larger number of points available. When the method of analysis given here is to be used, this can be done by first applying some process of graduation, to produce a smooth curve through each set of points corresponding to each animal, and analyzing the smoothed values read off from the curves, at a number of equal intervals sufficient to give an adequate description of the curves. This graduation of the data can sometimes be accomplished quite satisfactorily by fitting the curves by eye, but some may prefer a method which is more objective. Since any polynomial of degree p is uniquely determined by specifying $p + 1$ points through which it passes, the division of the smoothed curve into p periods, specified by $p + 1$ points, is equivalent to the description of the curve by a polynomial of degree p. In the growth experiment described here, two weighings were made per week, and the values actually plotted in Fig. II and analyzed, are the means of these pairs of values.

6. ELIMINATION OF THE INITIAL WEIGHT

In the analysis of growth curves the increases in weight of the animals may be correlated with their initial weights. In this case greater precision may be obtained if the analysis is made after elimination of the initial weight by covariance analysis (see for example Fisher 1941). The elimination of y_0, the initial weight, can be accomplished in a precisely similar manner to that used for the elimination of $\bar{y}$ from the criterion for "shape" analysis of 4.3, that is to say we have for the *overall* criterion after the elimination of y_0

$$\Lambda_{1234.0} = \Lambda_{01234}/\Lambda_0$$

This criterion serves to assess the differences in growth rates during the four periods, after the regression of each of these variates on the initial weight has been allowed for, and its significance is assessed as before, the degrees of freedom for error being one less than for the corresponding criterion Λ_{1234}. We shall normally wish to analyze $\Lambda_{1234.0}$ further, however, and to do this we shall need the corresponding matrices of sums of squares and products. These can be found in the manner described in §4.2. The matrices for error and for error plus treatments for the five variates y_0, y_1, y_2, y_3 and y_4 are reduced by a single pivotal condensation based on the element corresponding to the sum of squares for y_0. This gives the desired matrices for the numerator and denominator of $\Lambda_{1234.0}$. Further analysis of the data into differences in mean growth rate and differences in "shape" can be accomplished by operating on the determinants of $\Lambda_{1234.0}$ in precisely the same manner as has already been described for Λ_{1234}.

The two components $\Lambda(\bar{y} : y_0)$ and $\Lambda(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y} : y_0)$ assess respectively, the change in overall growth rate when change in initial weight is allowed for, and the change in shape when change in initial weight is allowed for; again apart from the loss of a degree of freedom the tests are the same.

Finally $\Lambda(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y} : y_0 , \bar{y})$ can be calculated by eliminating $\bar{y}$ in a precisely similar way as before and this criterion will assess whether group differences occur other than can be explained by the relation within the groups between the "shape" and the initial weight and average growth rate.

7. TESTING AN ASSUMPTION IN THE MULTIVARIATE ANALYSIS

Just as in a single-variate analysis of variance the assumption is usually made that the observations are normally distributed about their population mean values with *constant variance* so an analogous assumption that the variates are multinormally distributed about their mean values with constant variance-covariance matrix is made in the multivariate analysis of variance of §4.1. If the variance-covariance pattern changes markedly from group to group, this test may be invalidated. Also in the test of §3.1 concerning the form of the variance-covariance matrix, the sums of squares and products are pooled on the tacit assumption that the variances and covariances do not change from one treatment group to the next. If this were not so an averaging effect might occur so that even though individual groups showed departure from the simple set-up, the overall criterion computed from the pooled error sums of squares and products for all the groups, might show no such departure.

To test the assumption that the matrix of variances and covariances remains constant from one treatment group to the next, the present author (Box 1949) has employed the multivariate analogue of Bartlett's (1937) criterion which is used to test for constancy of variance in the univariate case.

We take as our criterion

$$M = N \log_e | s_{ij} | - \sum_l (\nu_l \log_e | s_{ijl} |)$$

where s_{ijl} is the usual unbiased estimate of variance or covariance between the i-th and j-th variable in the l-th sample based on sums of squares and products having ν_l degrees of freedom and there are k such samples, s_{ij} is the *average* variance or covariance $(\sum_l \nu_l s_{ijl})/N$ and $N = \sum_l \nu_l$ the total of the degrees of freedom. It will be noted that, as usual, the *determinants* of variances and covariances replace the single variances of the univariate criterion. It is perhaps worth noting that

again this criterion is invariant under linear transformation of the data, that is to say data which show lack of constancy in variance and covariance cannot be made more homogeneous by linear transformation.

The test will be illustrated for the growth data of rats set out in Table D for the variates y_1, y_2, y_3 and y_4, recording increases in weight in four successive weekly periods.

The individual matrices for sums of squares and products are as follows:

GROUP I

$$\begin{bmatrix} 210.5 & 13.5 & -7.5 & -13.5 \\ 13.5 & 202.5 & 224.5 & 110.5 \\ -7.5 & 224.5 & 310.9 & 117.5 \\ -13.5 & 110.5 & 117.5 & 258.5 \end{bmatrix}$$

$\nu_1 = 9$

GROUP II

$$\begin{bmatrix} 111.4 & 83.0 & 78.4 & 39.7 \\ 83.0 & 246.0 & 292.0 & 157.0 \\ 78.4 & 292.0 & 473.4 & 264.7 \\ 39.7 & 157.0 & 264.7 & 174.9 \end{bmatrix}$$

$\nu_2 = 6$

GROUP III

$$\begin{bmatrix} 260.4 & -54.0 & -126.4 & -100.8 \\ -54.0 & 160.5 & 110.0 & 77.0 \\ -126.4 & 110.0 & 262.4 & 76.8 \\ -100.8 & 77.0 & 76.8 & 419.6 \end{bmatrix}$$

$\nu_3 = 9$

Since $| s_{ijl} | = | c_{ijl} | / \nu_l^p$ the determinants of the variance-covariance matrices can be obtained directly from the determinants for the sums of squares and products and we find

$$\log_e | s_{ij1} | = 11.2700, \log_e | s_{ij2} | = 10.8357, \log_e | s_{ij3} | = 12.7008$$

the determinant for the average variances and covariances is found in a similar way from the total sums of squares and products matrix (9);

$$\log_e | s_{ij} | = 12.4473$$

and $M = 24 \times 12.4473 - 9 \times 11.2700 - 6 \times 10.8357 - 9 \times 12.7008 = 17.9838$. This logarithmic statistic M is a further example of the class discussed in §3.2, and approximations have been derived using the general theory referred to before.

For the χ^2 approximation the following quantities are calculated

$$A_1 = \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_i \frac{1}{\nu_i} - \frac{1}{N} \right) \qquad f_1 = \frac{1}{2}(k-1)p(p+1)$$

and $(1 - A_1)M$ is distributed as χ^2 with f_1 degrees of freedom. As before, a more precise approximation, which is useful when some of the degrees of freedom ν_i are small or p and/or k are large can be obtained using tables of F. We calculate

$$A_2 = \frac{(p-1)(p+2)}{6(k-1)} \left(\sum_i \frac{1}{\nu_i^2} - \frac{1}{N^2} \right)$$

and refer M/b to tables of F with f_1 and f_2 degrees of freedom where

$$f_2 = \frac{f_1 + 2}{A_2 - A_1^2} \qquad \text{and} \qquad b = \frac{f_1}{1 - A_1 - f_1/f_2}$$

In this particular example $A_1 = 0.2408$, $f_1 = 20$, $(1 - A_1)M = 13.5$ is therefore referred to tables of χ^2 with 20 degrees of freedom. The probability for the occurrence of a value as great or greater than this, when the variances and covariances are in fact constant from one group to the next, is thus about 0.85, and there is therefore no reason to doubt the homogeneity of the data in this respect.

This paper originated partly as the result of a note published by O. L. Davies (1947) criticising a method of analysis for growth curves proposed by W. S. Weil (1947).

I am indebted to Dr. Davies for proposing this problem, and to those of my colleagues who were responsible for the investigations which are mentioned. In conclusion I wish to warmly acknowledge the help and guidance I have received from Dr. H. O. Hartley in this work.

SUMMARY

In the analysis of growth and wear curves, the effects can often be simply interpreted by differencing the original data; these differences correspond to the average growth rates during successive periods. If the successive periods are treated as the level of a further factor "periods", the effect of treatments on mean rate is measured by the variation in the period averages and on the "shape" of the rate curve by the interaction of these treatments with "periods". The taking of differences sometimes results, at least approximately, in a very simple covariance pattern for the errors, and the analysis can then be made by a simple application of the technique of the analysis of variance. A test is given which makes it possible to decide whether this simple set-up is contradicted by the data. When the simple set-up is not appropriate, a multivariate extension of

the analysis of variance is used to make the tests. Certain simple properties of the criterion are discussed which facilitate the analysis and the elimination of variables such as initial weight. Finally, it is shown how an important assumption made in the multivariate analysis may be tested.

REFERENCES

Aitken, A. C. *Proc. Roy. Soc. Edin., 55*, 42, 1935.
Aitken, A. C. *Determinants and Matrices*, 5th ed., Edinburgh: Oliver and Boyd, 1948.
Bartlett, M. S. *Proc. Camb. Phil. Soc., 30*, 327, 1934.
Bartlett, M. S. *Proc. Roy. Soc. A., 160*, 268, 1937.
Bartlett, M. S. *Proc. Camb. Phil. Soc., 34*, 33, 1938.
Box, G. E. P. *Biometrika, 36*, 317, 1949.
Buist, J. M., Newton, R. G., and Thornley, E. R. *Trans. I. R. I.* (In the press).
Daniels, H. E. *Proc. Camb. Phil. Soc., 34*, 321, 1938.
Dwyer, P. S. *J. Am. Statist. Ass., 37*, 441, 1942.
Fisher, R. A. *Statistical Methods for Research Workers*, 8th. Ed., 1941. Edinburgh: Oliver and Boyd.
Kolodziejczyk, S. *Biometrika, 27*, 161, 1935.
Neyman, J. and Pearson, E. S. *Biometrika, 20A*, 175 and 263, 1928.
Pearson, E. S. and Wilks, S. S. *Biometrika, 25*, 353, 1933.
Pearson, K. (Editor) Tables of the Incomplete Beta Function. London: Biometrika 1934.
Penrose, L. S. *Ann. Eugen., London, 13*, 228, 1947.
Smith, C. A. B. *Ann. Eugen., London, 13*, 272, 1947.
Thompson, C. M. *Biometrika, 32*, 168, 1941.
Thompson, C. M. and Merrington, M. *Biometrika, 33*, 74, 1943.
Wilks, S. S. *Biometrika, 24*, 471, 1932.
Wilks, S. S. *Ann. Math. Statist., 17*, 257, 1946.
Wishart, J. *Biometrika, 30*, 16, 1938.
Wishart, J. *J. R. Statist. Soc. Suppl., 6*, 1, 1939.

TABLE D

INITIAL WEIGHT AND WEEKLY GAINS IN WEIGHT FOR 27 RATS

Group 1. Control						Group 2. Thyroxin						Group 3. Thiouracil					
Rat	y_0	y_1	y_2	y_3	y_4	Rat	y_0	y_1	y_2	y_3	y_4	Rat	y_0	y_1	y_2	y_3	y_4
1	57	29	28	25	33	11	59	26	36	35	35	18	61	25	23	11	9
2	60	33	30	23	31	12	54	17	19	20	28	19	59	21	21	10	11
3	52	25	34	33	41	13	56	19	33	43	38	20	53	26	21	6	27
4	49	18	33	29	35	14	59	26	31	32	29	21	59	29	12	11	11
5	56	25	23	17	30	15	57	15	25	23	24	22	51	24	26	22	17
6	46	24	32	29	22	16	52	21	24	19	24	23	51	24	17	8	19
7	51	20	23	16	31	17	52	18	35	33	33	24	56	22	17	8	5
8	63	28	21	18	24							25	58	11	24	21	24
9	49	18	23	22	28							26	46	15	17	12	17
10	57	25	28	29	30							27	53	19	17	15	18

y_0 represents initial weight of rat
y_1 gain in 1st week
y_2 gain in 2nd week
y_3 gain in 3rd week
y_4 gain in 4th week

THE RELATIVE FREQUENCY OF SPARSE CELL ELEMENTS—AN APPLICATION TO RETICULOCYTE BLOOD COUNTS

MARVIN SCHNEIDERMAN AND GEORGE BRECHER

National Institutes of Health, Public Health Service, Bethesda, Maryland

THE PROBLEM of determining the relative frequency of sparse elements frequently arises in making blood counts. For example, reticulocytes in normal humans comprise about 1% of the total red cells in the blood. With certain types of anemia this percentage rises considerably. The current, "standard" method for making reticulocyte counts involves counting 1000 red cells, and noting the number of reticulocytes among the 1000. This is a relatively time-consuming technique, and to reduce the work involved another technique has been suggested.

The new technique, which we call the Miller technique, involves using an ocular disc which is divided into two areas, as in figure 1, below.

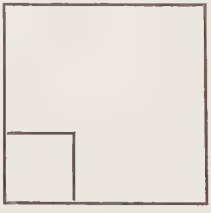


FIGURE 1
Miller ocular disc, ratio of large area to small area = 9.

The sparse elements are counted in the large area (the whole square) and the major elements (including sparse elements) are counted in the small area (the little square). The estimate of p (the proportion of the sparse elements) is then:

$$\hat{p} = \frac{y}{kx}$$

where y = the number of sparse elements counted
x = the number of major elements counted in the small area
k = the ratio of the large area to the small area

The question that then arises is whether this estimate of p can be made from the Miller technique with less work (for the same accuracy) than for the standard technique. It is the purpose of this note to show under what conditions the Miller ocular disc is a labor saving device.

It is common experience that in blood-element counting problems, the various elements appear to be distributed somewhat in accord with the Poisson law. We find that this accord is not perfect [1], [2], and, in fact, the internal variance may be less than Poisson. However, this difference is small, and is in the direction of conservatism. The difference from Poisson, if any, will tend to make firmer our estimate of p, using the Miller disc.

The variance of the estimated proportion, $\hat{p}$, is given to a first approximation by:

$$V(\hat{p}) = \frac{Y^2}{k^2X^2}\left(\frac{\sigma_y^2}{Y^2} + \frac{\sigma_x^2}{X^2} - \frac{2r_{xy}\sigma_x\sigma_y}{XY}\right) \tag{1}$$

where Y and X are the population values for which y and x are sample values.

Making the assumption of a Poisson distribution permits us to replace σ_y^2 by Y and σ_x^2 by X. If, in addition, we assume a zero correlation,* and assume that we use the Miller disc, with $k = 9$, we get:

$$V(\hat{p}) = \frac{Y}{81X^2}\left(1 + \frac{Y}{X}\right) \tag{2}$$

and since

$$p = \frac{Y}{9X},$$

$$V(\hat{p}) = \frac{p}{9X}(1 + 9p), \tag{3}$$

and the expected number of major elements, X, to be counted equals

$$X = \frac{p}{9V(\hat{p})}(1 + 9p) \tag{4}$$

*In using the Miller disc ($k = 9$), experimentally for making reticulocyte counts, a small positive, but unimportant, correlation was found. Neglecting the correlation will tend to overstate the variance. This error, like our assumption of the Poisson distribution, is in the direction of conservatism. The correlation becomes more important for values of k closer to unity.

To find where the Miller disc is useful, we want the total work of counting to be less for the Miller disc than for the standard technique. We can set up total work

$$W = X + cY \tag{5}$$

where: X = the number of major elements to be counted
Y = the number of sparse elements to be counted
c = the difficulty of counting a sparse element relative to the difficulty of counting a major element

so that from $p = Y/9X$ and (4) we get

$$W = \frac{p}{V(\hat{p})}\left(\frac{1}{9} + pc + p + 9cp^2\right) \tag{6}$$

Where the elements are of equal difficulty in counting $c = 1$ and

$$W = \frac{p}{9V(\hat{p})}(1 + 9p)^2 \tag{7}$$

In the standard technique, the variance of $\hat{p}$ equals $p(1 - p)/1000$, the usual binomial variation. Substituting this for $V(\hat{p})$ we have

$$W = \frac{1000}{9(1 - p)}(1 + 9p)^2 \tag{8}$$

Since our total work for the standard technique equals $1000(1 + p)$** cells counted, we may substitute $1000(1 + p)$ for W. Multiplying both sides of (8) by $9(1 - p)/1000$ we get

$$9(1 - p^2) = (1 + 9p)^2 \tag{8a}$$

and we find that for values of p less than .214 the Miller ocular is more efficient than the standard technique.

The relative amount of work for different values of p is given in column (2) in Table I.

The Miller disc is only one of many possible designs. At present the Miller disc is available commercially, but there may be more efficient designs. To investigate this we should return to equation (6). Replacing the 9's by k we have

$$W = \frac{p}{V(\hat{p})}\left(\frac{1}{k} + pc + p + ckp^2\right) \tag{9}$$

**For the standard technique the total work W may be considered equal to 1000 or $1000(1 + p)$ according to the mechanical process used in counting. For $W = 1000$ the Miller ocular is more efficient for p less than .189 rather than .214.

TABLE I
RELATIVE WORK USING THE MILLER TECHNIQUE

p	Relative Work	
	Miller disc, $k = 9$	Variable ocular, $k = 1/p$
(1)	(2)	(3)
.01	.132	.040
.02	.155	.080
.05	.233	.201
.10	.405	.404
.15	.628	.614
.20	.907	.833

The relative work multiplied by $1000(1 + p)$ gives the total number of cells to count using the Miller technique.

To minimize the total work, we differentiate with respect to k, and set the result equal to zero. We find, for a fixed value of $V(\hat{p})$, the value of k which minimizes the total work is:

$$k = \frac{1}{p\sqrt{c}} \tag{10}$$

where the two elements are of equal difficulty in counting $c = 1$ and $k = 1/p$. For this arrangement the same number of major and sparse elements will be counted.

Equation (8a) now becomes

$$\frac{1}{p}(1 - p^2) = 4 \tag{11}$$

and for $p < .236$ the Miller technique (with a variable ocular adjusted to $k = 1/p$) is more efficient than the standard technique. This value of p compares with the value of .214 obtained as the "efficiency" point when $k = 9$.

The gain, then, which arises in using a completely variable ocular (the Ehrlich ocular, for example) is quite small except for small values of p (See column (3), Table I). In addition, variable oculars are difficult to set at an exact value of $k = 1/p$ (especially when p is not known)*, so

*A variable ocular can be approximated roughly by using a hemocytometer and counting a different number of squares for the major elements and for the sparse elements. K is then set with some accuracy, and it may be possible to get close to $1/p$.

that a fixed-ratio ocular disc of the type of the Miller disc can be used without much loss.

Because of the region in which the Miller disc is most useful, we have computed fiducial limits for values of p from .01 to .20. Graphs incorporating these tables, and a report of experimental results using the Miller disc will be published elsewhere.

REFERENCES

[1] Berkson, J., Magath, T. B., and Hurn, M.: The Error of Estimate of the Blood Cell Count as Made with the Hemocytometer, *Am. Journal of Physiology*, 128: 309–323, 1940.

[2] Berkson, J.: Some difficulties of Interpretation Encountered in the Application of the Chi-Square Test, *Journal of the American Statistical Association*, 33: 203, p. 535, 1938.

STATISTICS SUMMER SESSION 1951

The Institute of Statistics is to hold a special summer session June 11–July 19, 1951. It will be for graduate students in statistics and for research workers, special emphasis being given to statistics in chemistry. Several visiting professors will participate in the lecturing. For details write the Institute of Statistics, State College Station, Raleigh, North Carolina.

A BIOMETRIC STUDY OF THE EXCRETION OF CORTICOSTEROIDS IN CHILDREN IN RELATION TO AGE, HEIGHT AND WEIGHT

MILDRED A. NORVAL, M.D.

Division of Biometry and Medical Statistics,
and
Section on Pediatrics,
Mayo Clinic,

and

NANCY KING, M.D.

Fellow in Pediatrics,
Mayo Foundation,
Rochester, Minnesota

THIS STUDY is based on data made available in an investigation by King and Mason (Ref. 1) in which the urinary excretion of corticosteroids was determined for 87 normal children (28 boys, 59 girls) including newly born babies and children up to 16 years of age. On approximately the same day as the urine was collected for the determination of the corticosteroids, measurements of height and weight without clothes were made. Urine was collected during twenty-four hours in glass bottles without preservative. The parents and the children were instructed as to the necessity of accurate collection of the urine. They were questioned afterward and if there was any doubt as to the accuracy of the collection, the specimen was discarded. The method for the determination of the corticosteroids has been described by Corcoran and Page (Ref. 2). The values for the urinary corticosteroids are expressed as milligrams of 11-dehydrocorticosterone per twenty-four hours.

The mean corticosteroid output per twenty-four hours increases steadily with age (fig. 1). Since, however, during this period the weight and the height are both increasing, it is pertinent to inquire whether the increase of output is related to either or both of these, rather than to age per se. Talbot (Ref. 3) reported that if the output of corticosteroids is expressed in ratio to surface area, the surface area being measured by the formula of DuBois ($S.A. = 0.00718\ W^{0.425}\ H^{0.725}$), the output does not change with age. This finding we did not confirm in our data. The mean

output per twenty-four hours per square meter of body surface (DuBois) is shown plotted against age in figure 1. It is seen that there is a sharp

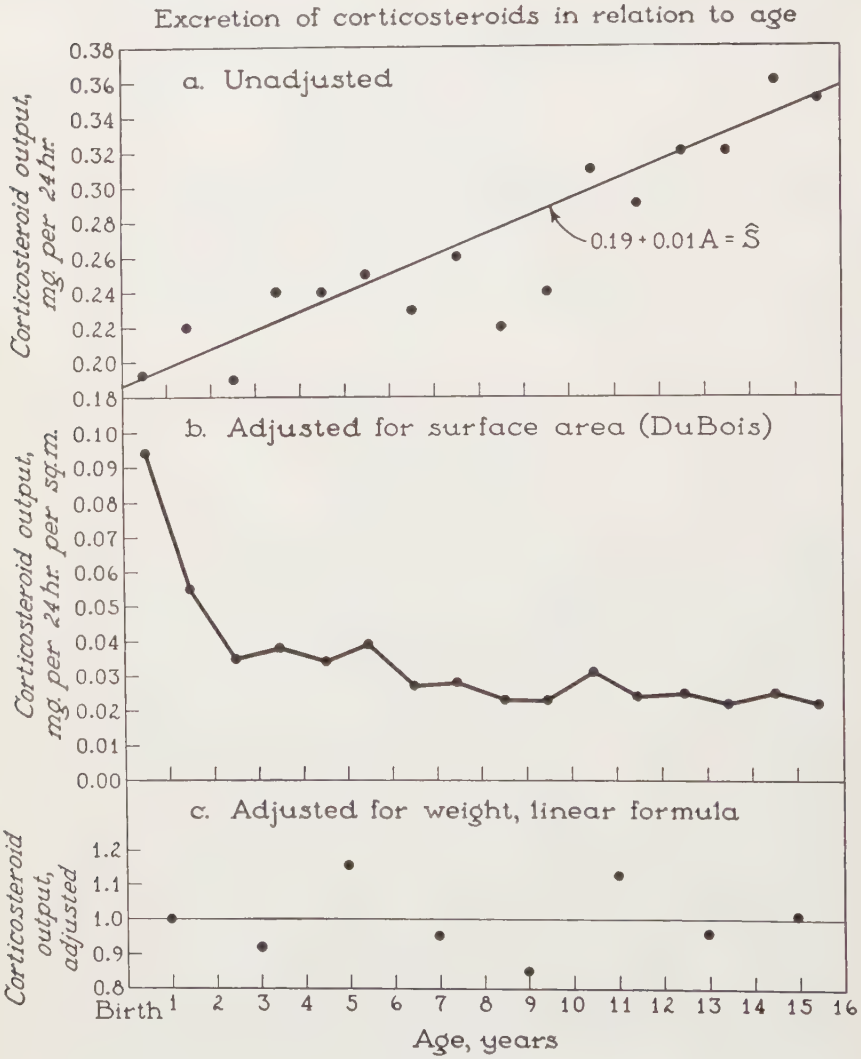


FIGURE 1

The relation of the mean corticosteroid output to age. a. Unadjusted. b. Adjusted for body surface (DuBois formula). c. Adjusted for weight, $S/\hat{S}$, $\hat{S} = 0.180 + 0.003\ W$, where S is the observed output, $\hat{S}$ is calculated from the linear formula, and W is the observed weight.

decrease in the first three years of age and a gradual decline thereafter.

However in this calculation we are of course not measuring surface

area directly, but utilize the exponential height-weight formula of DuBois with exponents $W^{0.425} H^{0.725}$. This amounts to standardizing corticosteroid output on the basis of an exponential formula with these particular exponents. It appeared possible that it is the inappropriateness of this height-weight formula for output of corticosteroids that accounts for the apparent downward regression with age. We therefore undertook to determine an estimating formula for output of corticosteroids in terms of weight and height obtained directly from the present data. Five formulas were evaluated for comparison:

$$\hat{S} = aW^{0.425} H^{0.725} \qquad \text{(DuBois)} \tag{1}$$

$$\hat{S} = aW^b H^c \tag{2}$$

$$\hat{S} = a\ W^b \tag{3}$$

$$\hat{S} = a + b\ W + c\ H \tag{4}$$

$$\hat{S} = a + b\ W \tag{5}$$

where $\hat{S}$ is the corticosteroid output estimated from the formula, W is weight in kilograms, H is height in centimeters, and a, b, c are parameters determined by fitting the equation in question by least squares.* The fit of the equations was compared by means of the estimated mean square deviation of the observed and predicted value of the output; the results are summarized in table 1. It is seen that the prediction using the exponents of the DuBois formula gives distinctly the poorest fit and that there is very little difference among the others. The simple linear equation in terms of weight alone gave the best fit as judged by the estimated variance, and this was therefore adopted. The correlation between the corticosteroid output and weight was +0.53. The corticosteroids adjusted for weight were then calculated for each individual as the ratio of the observed output S to the estimated normal output for weight, $\hat{S} = 0.180 + 0.003\ W$, (analogous with corticosteroids per unit "surface area"). The weight-adjusted corticosteroid output in relation to age is shown in figure 1. It is seen that there is no evident relation to

*It is common in fitting an exponential formula of the type $\hat{S} = a\ W^b$ to deal with the transformed linear equation $\log \hat{S} = a + b \log W$ and to fit the equation by minimizing $\Sigma(\log S - \log \hat{S})^2$ where $\hat{S}$ is the estimated and S the observed values. Since, however, in the linear least square solution we minimize $\Sigma(S - \hat{S})^2$ and since it is this sum which is used as the criterion of goodness of fit, it is this quantity that should be minimized for the exponential formula also. This may be accomplished by fitting the linear equation $\log \hat{S} = a + b \log W$ but using instead of $\log S$ with the actually observed S, a working value $S' = \log \hat{S}_1 + (S - \hat{S}_1)/\hat{S}_1$ where $\hat{S}_1$ is a value obtained from a preliminary fit, S is the observed output, and using for the weighting coefficient $W = \hat{S}_1^2$ (Ref. 4).

TABLE 1

PARAMETERS AND VARIANCE OF PREDICTION FORMULAS*

Prediction formula	Estimates			$\Sigma(S - \hat{S})^2$	D.F.	Estimated variance
	a	b	c			
$\hat{S} = a\ W^{0.425}\ H^{0.725}$	0.002			0.8088	86	0.00940
$\hat{S} = a\ W^b\ H^c$	10.473	0.163	0.291	0.5259	84	0.00626
$\hat{S} = a\ W^b$	6.920	0.287		0.5272	85	0.00620
$\hat{S} = a + b\ W + c\ H$	0.142	0.003	.0004	0.5251	84	0.00625
$\hat{S} = a + b\ W$	0.180	0.003		0.5156	85	0.00607

*$\hat{S}$ is the "predicted" output of corticosteroids (mg. per twenty-four hours), S the observed output. W is weight in kilograms, H is height in centimeters. The estimated values of a, b, c, were determined by least squares minimizing $\Sigma(S - \hat{S})^2$. The sum of squares $\Sigma(S - \hat{S})^2$ gives the actual sum of squared deviations of predicted and observed; this divided by the degrees of freedom (D. F.) gives an estimate of the variance about the hypothetical true curve rather than the fitted curve.

age. Therefore the essence of Talbot's conclusion that the amount of corticosteroids adjusted for body size is the same at all ages during the growth period is confirmed.

ACKNOWLEDGEMENT

We wish to thank Dr. J. Berkson for advice in the accomplishment of this study.

REFERENCES

1. King, Nancy and Mason, H. L.: Values of the Urinary Corticosteroids of Children. Unpublished data.
2. Corcoran, A. C. and Page, I. H.: Methods for the Chemical Determination of Corticosteroids in Urine and Plasma. *J. Lab. & Clin. Med.* 33:1326–1333, 1948.
3. Talbot, N. B.: Physiology of Adrenal Cortex. *Pediatrics.* 3:519–533, 1949.
4. Berkson, Joseph: Minimum X^2 and Maximum Likelihood Solution in Terms of a Linear Transform With Particular Reference to Bio-assay. *J. Amer. Statist. Assoc.* 44: 273–278, 1949.

THE EVALUATION OF DIAGNOSTIC TESTS

Samuel W. Greenhouse and Nathan Mantel

National Cancer Institute, Bethesda, Maryland

Introduction: The use of routine laboratory tests in diagnosing disease is becoming of increasing importance. This concentrated search for effective diagnostic tests emphasizes the need for procedures to evaluate and compare them. The following discussion presents such procedures (a) under the assumption of a continuous range of diagnostic test scores that are normally distributed in the universe and (b) where it is assumed that test scores are continuous but no assumption is made about the distribution forms.

Relatively few diagnostic tests correctly classify all persons tested as diseased or well. The more usual situation is one in which some well persons are classified as diseased and some diseased persons classified as well. Such errors do not necessarily invalidate a test, but they do make it necessary to define the maximum limits of error in an acceptable test. Thus, we can require, in general, that it classify correctly as diseased at least a minimum specified percentage, P_D , of those diseased, while classifying incorrectly as diseased no more than a specified percentage, P_W , of those well. For example, we might require that a test correctly classify at least 90% of the diseased while it incorrectly classifies no more than 5% of the well. The percentages specified for a given test must, of course, be determined from subject matter rather than statistical considerations. This raises the following general problems:

A diagnostic test for a disease is proposed and it is desired to test the hypothesis that it meets the specified requirements. A sample of known positives and known negatives is drawn; the diagnostic test is applied and the results are analyzed for consistency with the hypothesis. Then, how large a sample of positives and negatives is required if it is desired to test this hypothesis when the probabilities of Type I and Type II errors are set at α and β respectively? Secondly, given two diagnostic tests, which one is more effective in detecting disease?

For a diagnostic test which yields only two possible results, positive or negative, specification of α and β determines the sample size needed. In general, however, diagnostic tests yield a fairly continuous range of possible results, which are then broken down into regions which are labeled positive or negative. Frequently, intermediate regions are in-

serted. The choice of the point of demarcation of the positive-negative region cannot be independent of the two percents specified. One can always select a point of demarcation so that fewer than P_D percent of the diseased will be correctly classified or more than P_W percent of the well will be incorrectly classified. It would thus be possible to reject an acceptable test because of an improper selection of a critical point. In consequence, for any given set of experimental results designed to evaluate a diagnostic test there are at least two possible choices of a critical point for specified P_D and P_W : that which yields exactly P_D percent of the diseased correctly classified, or that which yields exactly P_W percent of the well incorrectly classified. This formulation follows directly from Berkson's [1] demonstration that there is no single omnibus measure, such as the biserial correlation coefficient, by which one can rate a test.[1] Thus, the problem considered here is to determine the inherent goodness of a procedure devised to detect disease. This is in contrast to the situation in which one desires to evaluate a diagnostic test for which a critical point has already been fixed.

In what follows, it is assumed that for the sample of well and diseased there are no assignable causes of variation in the diagnostic test values other than the distinction between well and diseased. In practice, this assumption may not be met and it may be necessary, unless the sample is drawn with this in mind, to subdivide the sample into rational subgroups, such as by age or sex. This paper does not take account of procedures treating this problem of heterogeneity.

EVALUATION OF A SINGLE DIAGNOSTIC TEST

The significance test. Let us say that a good test, for a particular disease, has been defined as one which detects 90% or more of the diseased while misclassifying no more than 5% of the well, and let us assume, without loss of generality, that the diagnostic test yields higher values for the diseased than for the well. Then, if the diagnostic test value exceeded by only 5% of the well is designated by W_5 (the 95th percentile of the well distribution) and the value exceeded by 90% of the diseased is designated by D_{90} (the 10th percentile of the diseased distribution), the hypothesis we wish to test is that $W_5 \leq D_{90}$ or $D_{90} - W_5 \geq 0$. If this hypothesis is false the diagnostic test is not a good one.

We assume here that the distribution of diagnostic test values for both well individuals and diseased individuals is normal, or can be normalized by some transformation but not necessarily with the same

[1] Neyman [2], in his work on diagnosis, has been concerned with a different problem: one in which it is not known which individuals examined are diseased and which are well.

variances. Then, if the mean and standard deviation for the well are designated by m_W and σ_W and for the diseased by m_D and σ_D,

$$W_5 = m_W + 1.645\,\sigma_W$$

$$D_{90} = m_D - 1.282\,\sigma_D \quad [2]$$

If we let $\Delta = D_{90} - W_5$ our hypothesis becomes

$$\Delta = m_D - 1.282\,\sigma_D - m_W - 1.645\,\sigma_W \geq 0$$

For any sample, we obtain estimates of these four parameters and may compute an estimate of Δ, say $\hat{\Delta}$, as

$$\hat{\Delta} = \overline{X}_D - 1.282\,s_D - \overline{X}_W - 1.645\,s_W \tag{1}$$

where s is the sample standard deviation

$$s_X = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}} \quad [3]$$

The estimate, $\hat{\Delta}$, is approximately normally distributed [3] even for moderately sized samples, so that if the variance were known exactly a test of the hypothesis $\Delta = 0$ is possible. Since all four variables in (1) are independent the variance of $\hat{\Delta}$ is

$$V(\hat{\Delta}) \cong \frac{\sigma_D^2}{n_D}\left(1 + \frac{(1.282)^2}{2}\right) + \frac{\sigma_W^2}{n_W}\left(1 + \frac{(1.645)^2}{2}\right) \tag{2}$$

where n_D and n_W are the respective sample sizes of diseased and well individuals. An estimate of V is then given by

$$\hat{V}(\hat{\Delta}) = \frac{s_D^2}{n_D}\left(1 + \frac{(1.282)^2}{2}\right) + \frac{s_W^2}{n_W}\left(1 + \frac{(1.645)^2}{2}\right) \tag{3}$$

It is now necessary only to compute the value of $\hat{\Delta}/\sqrt{\hat{V}(\hat{\Delta})}$ and to compare this value with the appropriate one-sided critical level. Though no exact distribution of this is available, for large size samples the normal distribution can be used, the critical levels being -1.645 for 5% and -2.326 for 1%.

---

[2] The values $+1.645$ and -1.282 correspond to the 5% and 90% points of the normal distribution. Similar values can be used for other percentage points.

[3] This definition of the sample standard deviation leads to expressions for the variances of $\hat{\Delta}$ and other sample statistics involving the factor $n/(n-1)$. For simplicity of presentation and computation, the large sample assumption that this factor is equal to unity has been made. The appendix presents the development in which factors are introduced to yield unbiased estimates of the various sample statistics and their variances.

Determination of sample size. In testing the hypothesis $\hat{\Delta} \geq 0$ one must decide in advance the maximum frequency of error in testing the hypothesis. The hypothesis can be tested using a small sample, but only with a great risk of accepting it when false. If the diagnostic test is capable of detecting only, say, 80% of the diseased at a cost of misclassifying 5% of the well, we might wish to be relatively sure of rejecting the hypothesis. If this alternative hypothesis is true, $D_{80} = W_5$, and the expected value of $\hat{\Delta}$ is $D_{90} - W_5 = D_{90} - D_{80} = -.44\sigma_D$, rather than zero. The actual value of $\hat{\Delta}$, however, will be less than its expected value plus $1.645\sigma_{\hat{\Delta}}$ 95% of the time. Thus in order to have 95% assurance that we will not accept the hypothesis that $D_{90} = W_5$, when in fact $D_{80} = W_5$, we wish $-.44\sigma_D + 1.645\sigma_{\hat{\Delta}}$ to be significantly negative, that is to be less than $-1.645\sigma_{\hat{\Delta}}$. We require, therefore, that

$$-.44\ \sigma_D + 1.645\ \sigma_{\hat{\Delta}} \leq -1.645\ \sigma_{\hat{\Delta}} \tag{4}$$

whence

$$\sigma_{\hat{\Delta}} \leq \frac{.44}{3.29}\ \sigma_D\ . \tag{5}$$

To determine minimum sample size, we solve the equality

$$\sigma_{\hat{\Delta}} = \frac{.44}{3.29}\ \sigma_D = \sqrt{\frac{\sigma_D^2}{n_D}\left(1 + \frac{(1.282)^2}{2}\right) + \frac{\sigma_W^2}{n_W}\left(1 + \frac{(1.645)^2}{2}\right)} \tag{6}$$

Let $K = \sigma_D^2/\sigma_W^2$. Then

$$\frac{\sigma_{\hat{\Delta}}}{\sigma_D} = \frac{.44}{3.29} = \sqrt{\frac{1}{n_D}\left(1 + \frac{(1.282)^2}{2}\right) + \frac{1}{Kn_W}\left(1 + \frac{(1.645)^2}{2}\right)} \tag{7}$$

For any assumed value of K, any set of values n_D and n_W which satisfy the above equation will reject the hypothesis $\Delta \geq 0$, 95% of the time when $\Delta = -.44\sigma_D$. If there exist limitations on the number of known diseased individuals which can be obtained in evaluating the diagnostic test, we are still free to determine the number of well to be obtained. If the cost of evaluating a test can be represented as a function of the number of diseased and well individuals to be tested, then n_D and n_W can be so selected as to minimize the cost. As a simple case, consider the cost to be linear in total number of persons tested, $T = n_D + n_W$. We have to minimize $n_D + n_W$ while keeping

$$\frac{1}{n_D}\left(1 + \frac{(1.282)^2}{2}\right) + \frac{1}{Kn_W}\left(1 + \frac{(1.645)^2}{2}\right) = \frac{1.822}{n_D} + \frac{2.353}{Kn_W}$$

(the argument of the square root in (7)) fixed. The usual computation leads to the familiar form

$$\frac{n_D}{n_W} = \sqrt{\frac{1.822K}{2.353}} = .880\sqrt{K}$$

Substitution in (7) yields

$$n_D = 102 + \frac{116}{\sqrt{K}}$$

$$n_W = \frac{132}{K} + \frac{116}{\sqrt{K}}$$

The sample sizes required for certain assumed values of K are shown below.

$K = \sigma^2_D/\sigma^2_W$	Sample Size Required		
	Well n_W	Diseased n_D	Total T
1/9	1534	449	1983
1/4	758	334	1092
1/2	427	266	693
1	248	218	466
2	148	184	332
4	91	160	251
9	54	140	194

It is clearly seen that the larger the value of K, the smaller is the sample size which must be drawn to test the hypothesis $W_5 \leq D_{90}$, with assurance of rejection if in fact $W_5 = D_{80}$. However, this relation between sample size and K does not hold for alternative hypotheses of the type $D_{90} = W_8$ or $D_{90} = W_{10}$ etc. This arises from the fact that for this type of alternative hypothesis the expressions for n_D and n_W would have K and $\sqrt{K}$ in the numerator instead of the denominator.

Selection of a critical point for an accepted diagnostic test. For purposes of statistical testing it has been desirable to disregard the selection of a critical point. Once a test has been validated, it is necessary, for working purposes, to select a critical point. The fitting of normal distributions implied in testing the hypothesis of goodness of a diagnostic test is of aid also in selecting a critical point.

Assume a population with 1 percent diseased, and a diagnostic test just capable of detecting 90% of the diseased at a cost of misclassifying 5% of the well. Such a test will yield, for the population, $5\frac{1}{2}$ false positives for each true positive detected. In the neighborhood of the critical point, however, the test will yield a greater number of false positives for each true positive detected. The number of false positives per true positive at this point may be so large that we may not wish to include it in the positive region, or it may be so small that we may wish to shift our critical point farther to the left.

If, for a population with proportion p diseased, we wish to include in the positive region, all points which yield r or fewer false positives per true positive, we need simply find the point at which the number of false positives equals r times the number of true positives. This is given by the solution for X to

$$\frac{(1-p)}{\sigma_W\sqrt{2\pi}}\exp\left\{-\frac{(X-m_W)^2}{2\sigma_W^2}\right\}=\frac{rp}{\sigma_D\sqrt{2\pi}}\exp\left\{-\frac{(X-m_D)^2}{2\sigma_D^2}\right\}$$

$$X=\frac{m_D\sigma_W^2-m_W\sigma_D^2}{\sigma_W^2-\sigma_D^2}\pm\frac{\sigma_W\sigma_D\sqrt{(m_W-m_D)^2+2(\sigma_W^2-\sigma_D^2)\log_e\frac{rp\sigma_W}{(1-p)\sigma_D}}}{\sigma_W^2-\sigma_D^2} \tag{8}$$

and for the case

$$\sigma_W=\sigma_D=\sigma$$

$$X=\frac{m_W+m_D}{2}+\frac{\sigma^2}{m_W-m_D}\log_e\frac{rp}{(1-p)} \tag{9}$$

For $\sigma_W \neq \sigma_D$, there are two solutions for X. Since we may not wish to trust the assumption of normality to the limit, we should ordinarily accept only one of these. It would be reasonable, however, if there were a great disparity of variances between the two distributions, but with means near each other, to use two critical values and to assign "wild" values in either direction to the same diagnosis. For σ_D very much greater than σ_W there would then be two positive regions and an intermediate negative region.

One possible critical point of interest is that for which $r = (1-p)/p$. At this point the number of false positives per true positive is the same as the ratio of the number well to the number diseased for the population. At this point the ordinates of the two unit area normal distributions are

equal, and the difference between the percentage true positives and and percentage false positives is a maximum. A disadvantage of the use of this point as a critical point is that in classifying persons as positive at this point we are doing only as well as we could do by a chance classification procedure. But it excludes from the critical region all points at which we would get results worse than chance, and therefore is the most extreme value we could consider.

The value of p, the population proportion diseased, is of interest also in the case where we do not have a group of known well individuals. Instead we may have only a group of known diseased, and a group of unknowns from which to sample. The value of p should be allowed for in determining the proportion of unknowns which may fall in the positive region.

COMPARISONS OF TWO DIAGNOSTIC TESTS

If we have two different diagnostic tests we may be interested in knowing which is the better test. Under some circumstances, the appropriate question may not be which is better, but rather, what is the best way of using both diagnostic tests simultaneously: setting up critical points for each test and requiring a positive value on one or both tests; or possibly setting up a linear combination of scores on the two tests which differentiates maximally between well and diseased. This question is related to the degree of independence of the two tests. In this paper attention will be restricted only to the question of which is the better diagnostic test.

Tests on independent samples. Diagnostic test 1 has been applied to one sample of known diseased and known well individuals, diagnostic test 2 to another sample and the results are to be used to test the hypothesis that the diagnostic tests are equally good. (Determination of sample sizes needed for testing this hypothesis is not considered in this section, but can be obtained by a method parallel to that shown previously.)

We cannot unequivocally state that one diagnostic test is better or worse than another, unless we specify the percentage of false positives, or false negatives, we are willing to accept. Making the assumption of normality, it is only when $\sigma_{D1}/\sigma_{W1} = \sigma_{D2}/\sigma_{W2}$ that a statement of the relative goodness of two diagnostic tests is independent of the acceptable percentage of false positives or false negatives. Otherwise one test may be superior for one choice of a critical point and inferior for another critical point. Any statistical procedure that does not take account of this is likely to prove seriously misleading in practice.

A more appropriate general procedure would appear to involve (a)

pre-selecting a maximum acceptable level of false positives, say 5%, (b) determining the critical points of the well distributions for this acceptable level for both tests, (c) computing the percentage of the diseased population which lies above this point for each test and (d) calling that test which yields a higher percentage above the critical point the better test *for the predesignated level of false positives.*

It is more convenient, however, to work with the normal deviate corresponding to the percentage. We shall thus define

$$G = \frac{1}{\sigma_D}(m_D - m_W - 1.645\,\sigma_W)$$

and the diagnostic test for which G is greater is the better test.

The sample estimate of G is given by

$$\hat{G} = \frac{1}{s_D}(\overline{X}_D - \overline{X}_W - 1.645\,s_W) \tag{10}$$

The variance of $\hat{G}$ is

$$V(\hat{G}) \cong \frac{G^2}{2n_D} + \frac{1}{n_D} + \frac{\sigma_W^2}{n_W\sigma_D^2} + \frac{(1.645\,\sigma_W)^2}{2n_W\sigma_D^2} \tag{11}$$

which can be estimated by

$$\hat{V}(\hat{G}) = \frac{\hat{G}^2}{2n_D} + \frac{1}{n_D} + \frac{s_W^2}{n_W s_D^2} + \frac{(1.645\,s_W)^2}{2n_W s_D^2} \tag{12}$$

The statistical test to be applied now requires estimating for each diagnostic test the value of G and its variance and computing

$$t = \frac{\hat{G}_1 - \hat{G}_2}{\sqrt{\hat{V}(\hat{G}_1) + \hat{V}(\hat{G}_2)}}$$

which can be considered a normal deviate and applying a two-sided significance test.

Tests on dependent samples. The statistical test just considered is appropriate when separate samples are drawn for the two diagnostic tests. If, however, the same individuals are used for both tests, we should expect to make more efficient use of our sample by taking into account the correlation of scores on the two diagnostic tests.

For this purpose we compute $\hat{G}_1 - \hat{G}_2$ in the same way as for independent samples. But now

$$V(\hat{G}_1 - \hat{G}_2) = V(\hat{G}_1) + V(\hat{G}_2) - 2CV(\hat{G}_1\hat{G}_2)$$

where CV indicates covariance and is given by

$$(13) \qquad CV(\hat{G}_1\hat{G}_2) \cong \frac{\rho_D^2 G_1 G_2}{2n_D} + \frac{\rho_D}{n_D} + \frac{\rho_W \sigma_{W1}\sigma_{W2}}{n_W \sigma_{D1}\sigma_{D2}} + \frac{(1.645\ \rho_W)^2 \sigma_{W1}\sigma_{W2}}{2n_W \sigma_{D1}\sigma_{D2}}$$

where ρ_D is the correlation of scores on the two diagnostic tests for the population of diseased individuals, ρ_W for the population of well individuals. Again, inserting sample values for the population parameters in (13), we obtain an estimate of the covariance, which together with the estimates of the variances of $\hat{G}_1$ and $\hat{G}_2$ provide the following estimate of $V(\hat{G}_1 - \hat{G}_2)$,

$$\hat{V}(\hat{G}_1 - \hat{G}_2) = \frac{1}{2n_D}(\hat{G}_1^2 + \hat{G}_2^2 - 2r_D^2\hat{G}_1\hat{G}_2 - 4r_D + 4)$$

$$(14) \qquad + \frac{2 + (1.645)^2}{2n_W}\left[\frac{s_{W1}^2}{s_{D1}^2} + \frac{s_{W2}^2}{s_{D2}^2}\right] - \frac{s_{W1}s_{W2}}{n_W s_{D1}s_{D2}}[2r_W + (1.645r_W)^2]$$

Then

$$t = \frac{\hat{G}_1 - \hat{G}_2}{\sqrt{\hat{V}(\hat{G}_1 - \hat{G}_2)}}$$

DISTRIBUTION-FREE SIGNIFICANCE TESTS[4]

In the preceding sections the significance tests have been based on the assumption of normality of the diagnostic test scores, or some transformation of them. Opposed to the procedure of computing a percentile on the basis of sample estimates of the population parameters is the procedure of estimating a percentile by counting in the sample. The latter estimate can always be made, assuming large enough samples. However, where the population distribution form can be assumed to be normal, counting is less efficient. It is well known, for example, that for a normal population, using the median has an efficiency of 64% compared with the use of the sample mean in estimating the 50th percentile. For other percentiles, the relative efficiency of counting is even less. For example, in the 5th to 10th percentile range, the relative efficiency of counting is approximately 50% to 60% and at the 0.1 percentile, the relative efficiency decreases to 6%.

If, however, it is believed the population distribution form differs sufficiently from normality to invalidate the calculation of a percentile

[4]Tests for comparing percentage points in arbitrary, continuous distributions have been considered by Marshall and Walsh [4]. For large samples, the methods presented here are equivalent to those given in [4].

as $m + t\sigma$, an estimate of the population should be made by counting in the sample.[5]

Evaluation of a single test. If we estimate W_5 and D_{90} by counting in our sample the variances of the estimates are approximately

$$\frac{(.05)(.95)}{n_W z^2_{W_5}} \quad \text{and} \quad \frac{(.10)(.90)}{n_D z^2_{D_{90}}}$$

respectively. z_{W_5} and $z_{D_{90}}$ are the probability densities per unit interval in the respective distributions at W_5 and D_{90}. Estimates of z_{W_5} and $z_{D_{90}}$ can be made by smoothing the data of the samples and using the smoothed ordinate. The test of the hypothesis $W_5 \leq D_{90}$ is then given by

$$t = \frac{\hat{D}_{90} - \hat{W}_5}{\sqrt{\frac{(.05)(.95)}{n_W \hat{z}^2_{W_5}} + \frac{(.10)(.90)}{n_D \hat{z}^2_{D_{90}}}}} \tag{15}$$

Comparisons of two tests-independent samples. For comparison of two diagnostic tests, using independent samples, we must again specify the acceptable percentage of false positives or false negatives. For 5% false positives acceptable we consider that test better which gives the larger percentage of true positives. By counting, for diagnostic test 1, we find at $\hat{W}_5$ the percentage true positives in the sample of positives to be $\hat{P}_{D1}$. The variance of the estimate $\hat{P}_{D1}$, taking into account the variance of $\hat{W}_5$, is

$$V(\hat{P}_{D1}) \cong \frac{P_{D1}(1 - P_{D1})}{n_{D1}} + \frac{(.05)(.95)}{n_{W1}} \times \frac{z^2_{D1}}{z^2_{W1}} \tag{16}$$

where the z's are the probability densities at W_5 for the respective distributions, and can be estimated as before. $\hat{P}_{D1}$ can be used in place of P_{D1}, and for large enough n, the bias correction in $\hat{P}^2_{D_1}$, is trivial. If similar estimates are made for diagnostic test 2, the significance test becomes

$$t = \frac{(\hat{P}_{D1} - \hat{P}_{D2})}{\sqrt{\frac{\hat{P}_{D1}(1-\hat{P}_{D1})}{n_{D1}} + \frac{\hat{P}_{D2}(1-\hat{P}_{D2})}{n_{D2}} + (.05)(.95)\left[\frac{1}{n_{W1}}\left(\frac{\hat{z}_{D1}}{\hat{z}_{W1}}\right)^2 + \frac{1}{n_{W2}}\left(\frac{\hat{z}_{D2}}{\hat{z}_{W2}}\right)^2\right]}} \tag{17}$$

[5]John Tukey makes the valuable suggestion that non-normality may not seriously affect the approximation of the 5% point as $m + 1.645\sigma$, but the effect of non-normality on the variance of s should be taken into account in testing significance if we use procedures which assume normality. This can be done by multiplying $\sigma^2/2n$, the variance of s assuming normality, by $(\beta_{2-1})/2$, where $\beta_2 = \mu_4/\mu^2_2$, for which an estimate can be obtained from the data.

Dependent samples. Where the same sample is used for both diagnostic tests we may obtain $\hat{P}_{D1}$ and $\hat{P}_{D2}$ as above. The covariance of $\hat{P}_{D1}$, and $\hat{P}_{D2}$ must be taken into account however. These variances and covariance are composed of two independent parts—a part due to variation of diseased individuals, and a part due to the variation of the critical point.

Making the usual Taylor series approximations we obtain for the variance of $\hat{P}_{D1} - \hat{P}_{D2}$

$$V(\hat{P}_{D1} - \hat{P}_{D2}) \cong \frac{P_{D1}(1 - P_{D1})}{n_D} + \frac{P_{D2}(1 - P_{D2})}{n_D}$$

$$- \frac{2\rho'_D}{n_D}\sqrt{P_{D1}P_{D2}(1 - P_{D1})(1 - P_{D2})} \tag{18}$$

$$+ \frac{(.05)(.95)}{n_W}\left[\left(\frac{z_{D1}}{z_{W1}}\right)^2 + \left(\frac{z_{D2}}{z_{W2}}\right)^2 - 2\rho'_W \frac{z_{D1}}{z_{W1}}\frac{z_{D2}}{z_{W2}}\right]$$

where ρ'_D is the correlation of the diagnostic tests in the diseased population, ρ'_W in the well population, using scores of zero for values less than the population W_5, scores of one for values greater than W_5.

If in our sample of diseased individuals, a_D individuals are classified as positives (exceeding $\hat{W}_5$) by both tests, the sample value r'_D is then

$$r'_D = \left(\frac{a_D}{n_D} - \hat{P}_{D1}\hat{P}_{D2}\right) \Big/ \sqrt{\hat{P}_{D1}\hat{P}_{D2}(1 - \hat{P}_{D1})(1 - \hat{P}_{D2})} \tag{19}$$

And the sample value, r'_W, if a_W well individuals are classified as positive by both tests, is

$$r'_W = \left(\frac{a_W}{n_W} - (.05)^2\right) \Big/ (.05)(.95) \tag{20}$$

Then, for large enough samples, using sample values of the P's, z's, and ρ''s, we have,

$$\hat{V}(\hat{P}_{D1} - \hat{P}_{D2}) = \frac{1}{n_D}\left[\hat{P}_{D1}(1 - \hat{P}_{D1}) + \hat{P}_{D2}(1 - \hat{P}_{D2})\right.$$

$$\left. + 2\hat{P}_{D1}\hat{P}_{D2} - \frac{2a_D}{n_D}\right] + \frac{(.05)(.95)}{n_W}\left\{\left(\frac{\hat{z}_{D1}}{\hat{z}_{W1}}\right)^2 + \left(\frac{\hat{z}_{D2}}{\hat{z}_{W2}}\right)^2\right\} \tag{21}$$

$$- \frac{2\hat{z}_{D1}\hat{z}_{D2}}{n_W\hat{z}_{W1}\hat{z}_{W2}}\left(\frac{a_W}{n_W} - (.05)^2\right)$$

$$t = (\hat{P}_{D1} - \hat{P}_{D2}) / \sqrt{\hat{V}(\hat{P}_{D1} - \hat{P}_{D2})} \tag{22}$$

This test cannot be applied when we are considering two diagnostic tests which are both so good that there is no overlapping of values in the sample for diseased and well individuals. The test assuming normality, when justified, however, is applicable whether or not there is overlapping.

APPENDIX

For the development of estimates and variances below and in the text, use has been made of the following properties of the normal and bivariate normal distributions:

$$E(\bar{x}) = m$$

$$E(s^2) = \sigma^2$$

$$E(s) = \frac{\sigma \sqrt{\frac{2}{n-1}}\, \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cong \sigma \sqrt{\frac{n - 1\frac{1}{2}}{n-1}}$$

$$E\left(\frac{1}{s}\right) = \frac{1}{\sigma} \frac{\sqrt{\frac{n-1}{2}}\, \Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cong \frac{1}{\sigma} \sqrt{\frac{n-1}{n - 2\frac{1}{2}}}$$

$$E\left(\frac{1}{s^2}\right) = \frac{1}{\sigma^2} \frac{n-1}{n-3}$$

$$V(s) \cong \frac{\sigma^2}{2(n-1)}$$

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

$$\rho_{\bar{x},\bar{y}} = \rho_{x,y}$$

$$\rho_{s_x,s_y} = \rho^2_{x,y}$$

$$\rho_{\bar{x},s_x} = \rho_{\bar{x},s_y} = \rho_{\bar{y},s_x} = \rho_{\bar{y},s_y} = 0$$

Use is made of the Taylor series approximation for obtaining vari-

ances and covariances of products and ratios, with expansion about expected values, e.g.

$$V\left(\frac{1}{s}\right) \cong \left[E\left(\frac{1}{s}\right)\right]^4 V(s)$$

$$CV\left(\frac{1}{s_x}, \frac{1}{s_y}\right) \cong \left[E\left(\frac{1}{s_x}\right)E\left(\frac{1}{s_y}\right)\right]^2 CV(s_x, s_y)$$

$$V\left(\frac{\bar{x}}{s}\right) \cong \left(E\left(\frac{1}{s}\right)\right)^2 V(\bar{x}) + \left(E\left(\frac{1}{s}\right)\right)^4 (E(\bar{x}))^2 V(s)$$

The approximation to $V(1/s)$ instead of its exact value is made for consistency with the approximation used for $CV(1/s_x, 1/s_y)$.

We can then obtain the following formulas corresponding to those in the text giving unbiased estimates (the subscript u referring to unbiased).

$$\Delta = m_D - 1.282\,\sigma_D - m_W - 1.645\,\sigma_W$$

$$\hat{\Delta}_u = \bar{x}_D - 1.282\,s_D\sqrt{\frac{n_D - 1}{n_D - 1\frac{1}{2}}} - \bar{x}_W - 1.645\,s_W\sqrt{\frac{n_W - 1}{n_W - 1\frac{1}{2}}}$$

$$V(\hat{\Delta}_u) \cong \frac{\sigma_D^2}{n_D} + \frac{(1.282\,\sigma_D)^2}{2n_D - 3} + \frac{\sigma_W^2}{n_W} + \frac{(1.645\,\sigma_W)^2}{2n_W - 3}$$

$$\hat{V}(\hat{\Delta}_u)_u = \frac{s_D^2}{n_D}\left(1 + \frac{(1.282)^2\,n_D}{2n_D - 3}\right) + \frac{s_W^2}{n_W}\left(1 + \frac{(1.645)^2\,n_W}{2n_W - 3}\right)$$

For $$G = \frac{1}{\sigma_D}(m_D - m_W - 1.645\,\sigma_W)$$

$$\hat{G}_u = \frac{1}{s_D}\sqrt{\frac{n_D - 2\frac{1}{2}}{n_D - 1}}\left(\bar{x}_D - \bar{x}_W - 1.645\,s_W\sqrt{\frac{n_W - 1}{n_W - 1\frac{1}{2}}}\right)$$

$$V(\hat{G}_u) \cong \frac{G^2}{2n_D - 5} + \frac{1}{n_D} + \frac{\sigma_W^2}{n_W\sigma_D^2} + \frac{(1.645\,\sigma_W)^2}{(2n_W - 3)\sigma_D^2}$$

$$\hat{V}(\hat{G}_u)_u = \frac{2n_D - 6}{2n_D - 5}\left(\frac{\hat{G}_u^2}{2n_D - 5} + \frac{1}{n_D} + \frac{(n_D - 3)s_W^2}{n_W(n_D - 1)s_D^2} + \frac{(n_D - 3)(1.645\,s_W)^2}{(n_D - 1)(2n_W - 3)s_D^2}\right)$$

$$\widehat{(G_1 - G_2)}_u = \hat{G}_{1u} - \hat{G}_{2u}$$

$$V\widehat{(G_1 - G_2)}_u = V(\hat{G}_{1u}) + V(\hat{G}_{2u}) - 2CV(\hat{G}_{1u}, \hat{G}_{2u})$$

For independent samples $CV(\hat{G}_{1u}\ ,\ \hat{G}_{2u}) = 0$. For identical samples

$$CV(\hat{G}_{1u}\ ,\ \hat{G}_{2u}) \cong \frac{\rho_D^2 G_1 G_2}{2n_D - 5} + \frac{\rho_D}{n_D} + \frac{\rho_W \sigma_{W1}\sigma_{W2}}{n_W \sigma_{D1}\sigma_{D2}} + \frac{(1.645\ \rho_W)^2 \sigma_{W1}\sigma_{W2}}{(2n_W - 3)\sigma_{D1}\sigma_{D2}}$$

for which no unbiased estimate is obtained.

Although no tables exist for making the exact tests of significance implied above, for the large size samples necessary to evaluate diagnostic tests, the use of the normal deviate will be appropriate. For such samples, there will be only a trivial difference between the biased and unbiassed forms. However, it is likely that existing tables of t and the normal deviate come closer to describing the unbiased rather than the biased forms. In addition, to the extent that G may be considered a population parameter measuring the goodness of a diagnostic test, the unbiased estimate of G may be preferred.

REFERENCES

[1] Joseph Berkson—"Cost-Utility" as a Measure of the Efficiency of a Test, *Journal of the American Statistical Association* 42, 238, 246–255, 1947.

[2] J. Neyman—Outline of Statistical Treatment of the Problem of Diagnosis, *Public Health Reports*, 62, 40, 1947.

[3] Eisenhart, Hastay and Wallis—*Selected Techniques of Statistical Analysis*, Ch 1, New York, 1947.

[4] A. W. Marshall and J. E. Walsh—*Some Tests for Comparing Percentage Points of Two Arbitrary Continuous Populations*, The Rand Corporation, P-133, March 1950.

ESTIMATES OF THE LD_{50} : A CRITIQUE

Irwin Bross

*Department of Biostatistics**
The Johns Hopkins University

ABSTRACT

Karber's (Spearman's) Method, the Reed-Muench (Behren's) Method, and the Cornfield-Mantel iterative approximation to the maximum-likelihood estimate of the LD_{50} are compared. The performance of each method is evaluated by small sample analysis, and an index of "worth" is introduced to facilitate the comparison.

ESTIMATES OF THE LD_{50}

The simplest stimulus-response problem, and the only one considered here, employs an experimental design with several dosage levels of some deleterious substance and the same number of test animals at each level. The purpose of the experiment is to estimate the dosage level at which 50% of the animals would die. This dosage level, generally known as the LD_{50} , is then used as an index of the potency of the toxin.

The various estimates of the LD_{50} are relatively old, one of the most venerable being "Karber's Method" which was developed by Spearman in 1908 (1). The method has a curious history of popularity and disuse. It was employed by Karber for bioassay in 1930 but ran into criticism shortly thereafter. It was advocated (with reservations) by Irwin (2) but rejected by Finney (3). Cornfield and Mantel (4) discuss the theoretical derivation and use it as a first approximation to the maximum likelihood solution.

The Reed-Muench Method (5), which is also called Behren's Method, has been used by research workers for a quarter of a century but has also been heavily criticized. One objection, that no estimate of standard error was available, was recently overcome by Dr. Mario Pizzi in his doctoral dissertation (6).

Although the "probit" method is also an old one, it did not attain popularity until the middle thirties at which time the method of maximum

*Departmental Paper No. 265.

likelihood was applied and certain computational difficulties were overcome by tabulation. Because of its superior theoretical foundations, it has been strenuously advocated. The "probit" method (a) assumes that the probability of death at any level follows an integrated normal distribution, and (b) uses the method of maximum likelihood to obtain the estimate.

A catalogue of the variant methods will not be attempted here, but the divergences take two directions. Other fundamental curves have been proposed to replace the normal. The most important of these is the integrated logistic curve, which will be used in this analysis, but various other candidates are in the race. Principles of estimation other than maximum likelihood, such as least squares, minimum chi square (7), and mean moving averages (8), have also been suggested. Over twenty different estimates of the LD_{50} are currently on the market.

The partisans for the different estimates have advanced conflicting claims concerning the efficacy of their methods. The main arguments concern (1) computational convenience, (2) improved accuracy and generality, and (3) closer correspondence with the real world.

Insofar as the computations are concerned, the Spearman-Karber Method is the simplest, although the Reed-Muench Method is only slightly more difficult. The methods approximating the maximum likelihood solution are all fairly tedious.

The question of correspondence with the real world is one which must be considered separately for each experimental situation. However all of the fundamental curves currently employed are symmetric, unimodal, two-parameter curves.

It is in the consideration of accuracy that most of the controversy arises. One of the main reasons for this confusion is that the arguments are often based on large (infinite) sample statistical theory although the experimental designs often have twenty animals or less in the sample.

The purpose of this paper is to examine the question of accuracy in small samples.

Three methods of estimation were chosen for comparison. The Spearman-Karber and Reed-Muench Methods were selected because they are currently used and are well suited to the computational facilities of the ordinary research worker. Since the logistic was chosen as the fundamental curve, a direct comparison cannot be made with the Probit method. However Cornfield and Mantel have developed (4) an iterative approximation to the maximum likelihood solution for the logistic case which is analogous to the Probit method. This Cornfield-Mantel method was therefore chosen for comparison.

SMALL SAMPLE ANALYSIS

While methods of estimation may be shown to have very desirable properties in large samples, such as minimum variance, it does not follow that these advantages are retained when the sample size becomes small (and practical). In order to make a valid critique, it is therefore necessary to resort to small sample analysis—a complete enumeration of cases. Unfortunately this is a "brute-force" approach and requires extensive computation.

As a first step, the logistic was chosen as the fundamental response curve. The logistic resembles the normal curve in appearance, but it has somewhat longer tails. Moreover, the integral of the logistic has a simple form, and there is some evidence that the logistic provides a better "fit" in certain applications. It was especially convenient because a good deal of the necessary computation had already been done by Dr. Pizzi for this curve.

It would have been possible to fix the dosage levels and to vary the two parameters of the logistic, but it was somewhat easier (and mathematically equivalent) to work with the function.

$$(1.01) \qquad Pr(\text{an animal dies at dosage } d_k) = \frac{1}{1 + e^{-d_k}}$$

and to vary the location and spacings of the levels.

Thus it was assumed that

$$(1.02) \qquad d_k = d_1 + (k - 1)D$$

where d_1 is the dosage at the lowest level, and D is the spacing between levels.

It is common biological practice to use the logarithms of the dosages as an index in order to improve the fit with the fundamental curve. The actual doses would be in geometrical progressions and it would be the logarithm of the k-th dose which satisfied (1.02).

If we now consider an experimental design with m levels ($k = 1, 2, \cdots, m$) and n animals at each level, there will be $(n + 1)^m$ possible outcomes. The probability of any outcome may be found as a product of binomial probabilities.

For any choice of "spacing", D, and "location", d_1, the probability that an animal will die at a particular level may be found from (1.01). These values, substituted in the product of binomial probabilities, will give a numerical value for the probability of each outcome.

The next step is to calculate, for each outcome, the estimates of the LD_{50} by the three methods of estimation. This provides, in numerical form, the distributions of the estimates.

It will be noted that the amount of computation rapidly becomes astronomical. The special case $m = 4$ and $n = 2$ (81 outcomes) was evaluated for nine spacing-location configurations. The special case $m = 4$ and $n = 5$ (1296 outcomes) was evaluated for two configurations. It is realized that the conclusions afforded by this limited number of cases are not incontrovertible, and it is hoped that additional small sample analysis may fill some of the gaps.

A CRITERION FOR COMPARISON

It is not at once obvious what standards for accuracy can be used in small sample analysis. In large sample theory, the variance is the usual index for comparison. Three difficulties arise in the small sample case: (1) estimates may be infinite or may not exist for certain outcomes, (2) some outcomes are so peculiar (i.e. mortality decreases with dosage) that they might plausibly be rejected altogether, (3) the distributions may be markedly non-normal so that the "variance" is subject to misinterpretation. Due to the process of squaring, aberrant estimates may contribute heavily to the variance although their probabilities are small, and many workers would prefer to repeat such experiments.

In view of the objections, the index of comparison used in this critique is

$$W = Pr\,(\mid T - \theta \mid \leq \delta) \tag{1.03}$$

where T is the estimate, θ its true value, δ a positive number. Equation (1.03) states that W is the probability that the absolute error in the estimate is less than or equal to δ. A plot of W against δ was made for the different spacing-location configurations and for all three methods of estimation.

The choice of this criterion is necessarily somewhat arbitrary. A more precise treatment would consider the consequences of errors of various sizes. The criterion is essentially based on equal costs for error in each direction and an all-or-none cost function.

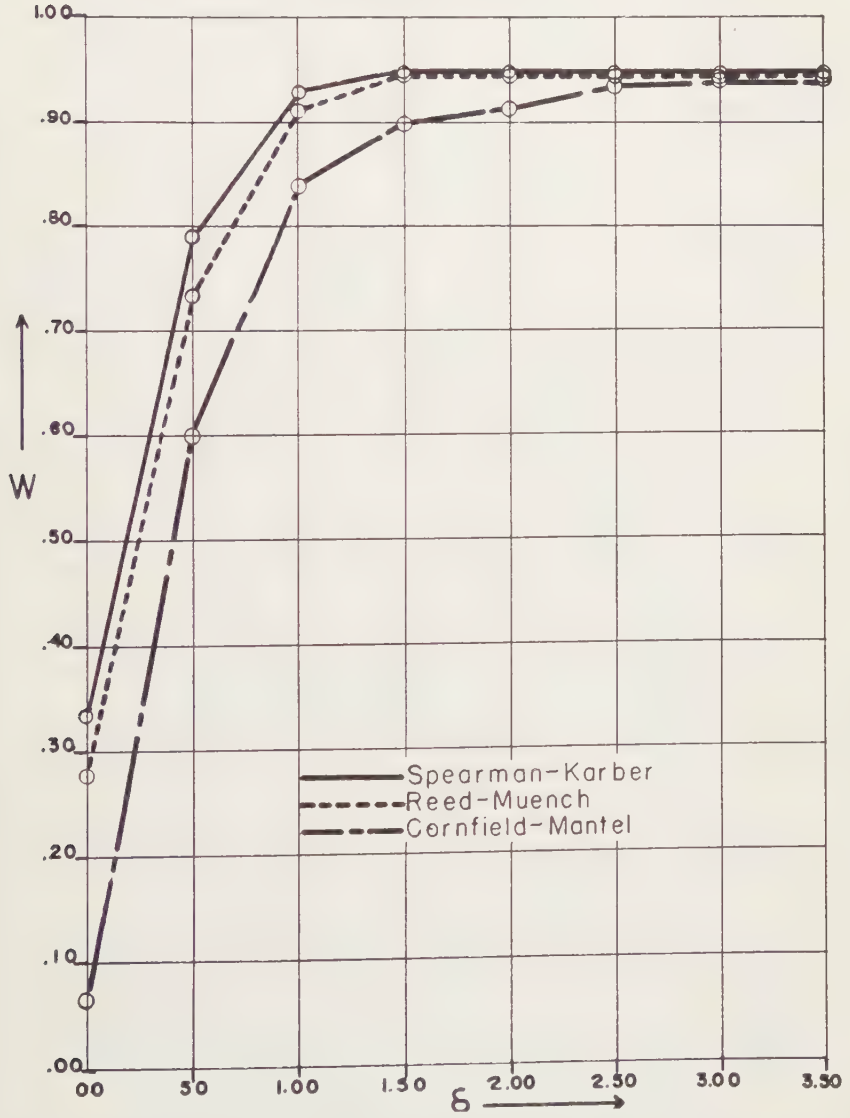
In other words, if we assume that an estimate is "worthwhile" if it is in the interval $\theta \pm \delta$, and "worthless" otherwise, then W would be an index of the "worth" of a method of estimation. Presumably the value a research worker would choose for δ would be different in different experimental situations. However if the worth of one method of estimation was greater than the worth of a second method for *all values of* δ, it would seem reasonable to say that the first method was superior.

TWO ANIMALS PER LEVEL

When four levels are used with two animals per level, the number of outcomes, 81, is small enough so that the case may be fairly thoroughly

GRAPH I

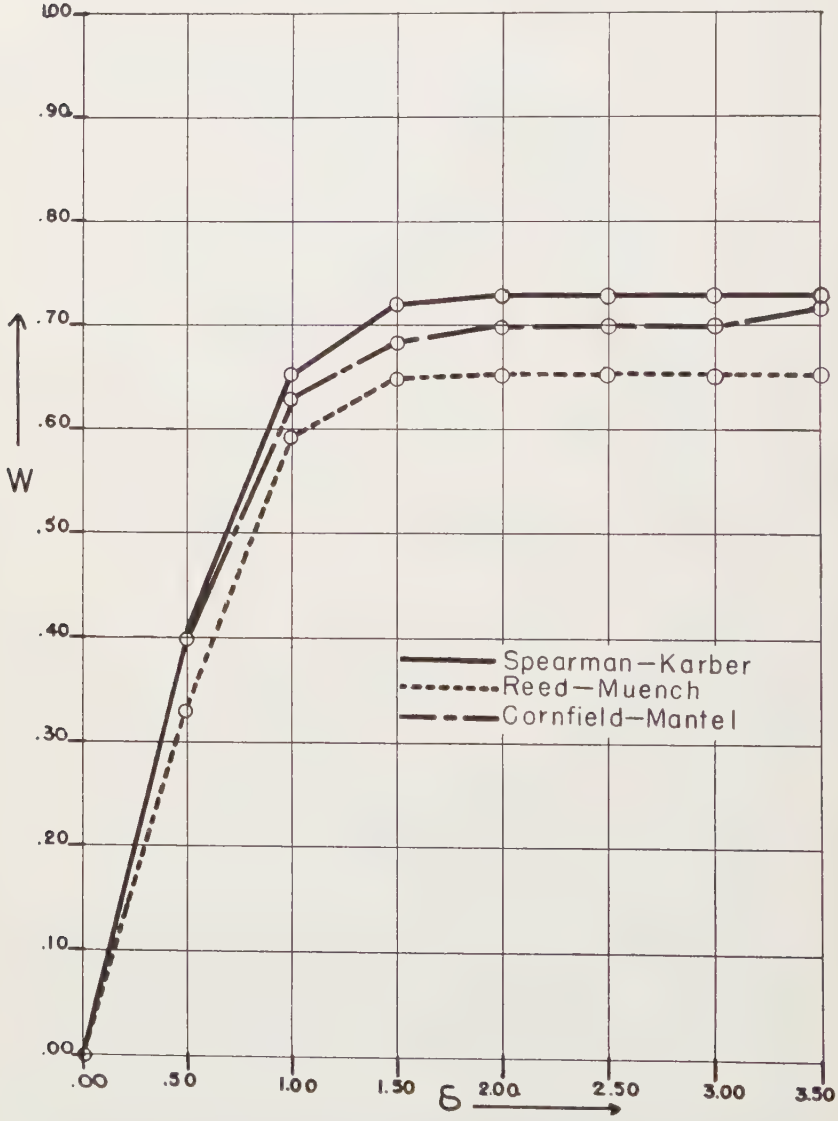
"WORTH" OF ESTIMATES—2 ANIMALS AT EACH OF 4 LEVELS WITH PROBABILITY OF DEATH AT LEVELS EQUAL TO .10, .32, .68 AND .90.
W = Probability ($| T - \theta | \leq \delta$)



explored. The Cornfield-Mantel (CM) estimate was calculated for thirty cases. The omitted cases were those with marked peculiarities. The cases calculated accounted for about 95% of the probability in all spacing-location configurations studied. The Spearman-Karber (SK) and Reed-Muench (RM) curves are for these same thirty cases.

GRAPH II

"WORTH" OF ESTIMATES—2 ANIMALS AT EACH OF 4 LEVELS WITH PROBABILITY OF DEATH AT LEVELS EQUAL TO .59, .86, .96 AND .99.
W = Probability ($| T - \theta | \leq \delta$)

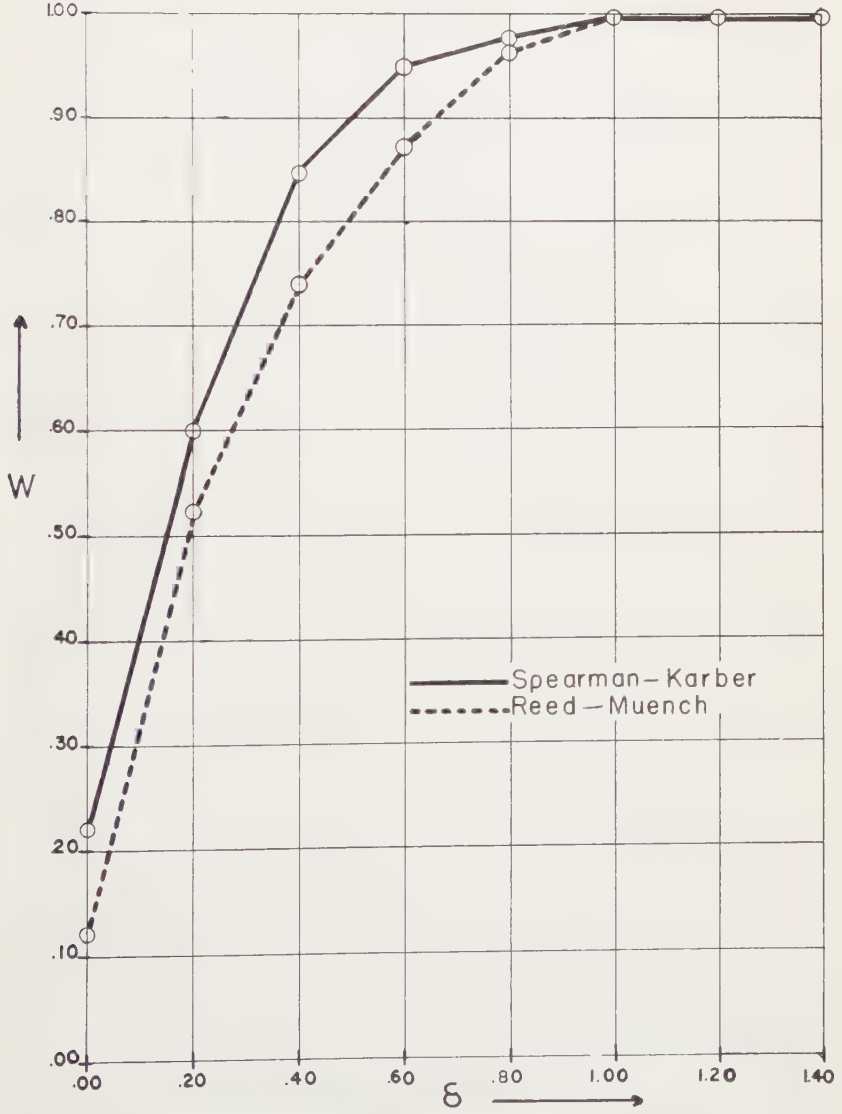


Graphs I and II may be regarded as typical of the nine configurations computed. Graph I is for the case where the probabilities of death at the four levels were .10, .32, .68, and .90 respectively and resembles the symmetric case for narrower spacings. In Graph II the corresponding probabilities are .59, .86, .96, and .99, so that the LD_{50} lies *outside* the levels used. For this case all the animals would die about a quarter of

GRAPH III

"WORTH" OF ESTIMATES—5 ANIMALS AT EACH OF 4 LEVELS WITH PROBABILITY OF DEATH AT LEVELS EQUAL TO .10, .32, .68 AND 90.

W = Probability ($| T - \theta | \leq \delta$)

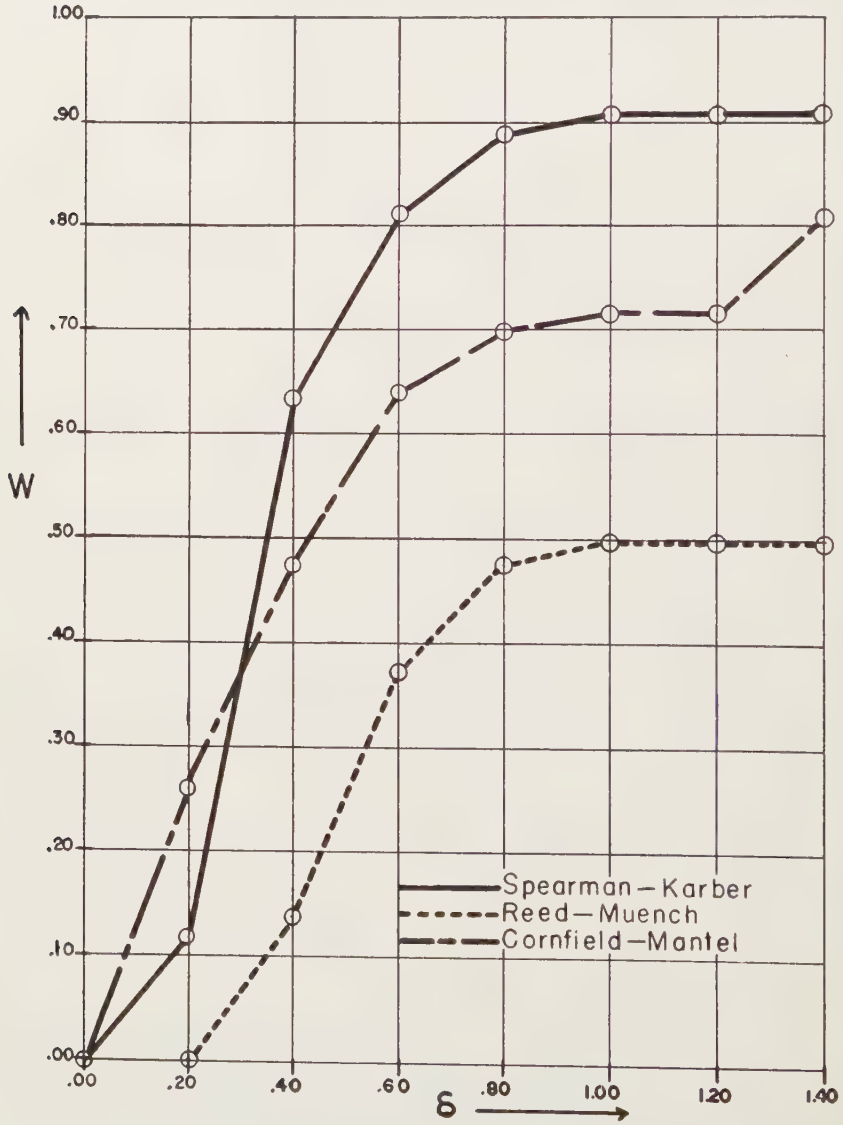


the time. There is no finite RM or CM estimate for this outcome. The cases of milder asymmetry had graphs more or less intermediate between Graphs I and II.

In nearly all cases W_{SK} was greater than either W_{RM} or W_{CM} for all values of δ. The single exception was one in which a CM was closer to θ than any SK estimate so that for a small interval near the origin,

GRAPH IV

"WORTH" OF ESTIMATES—5 ANIMALS AT EACH OF 4 LEVELS WITH PROBABILITY OF DEATH AT LEVELS EQUAL TO .59, .86, .96, AND .99.
W = Probability ($| T - \theta | \leq \delta$)



W_{CM} was higher. The advantage was quickly lost as δ increased. The relationship between W_{RM} and W_{CM} was not consistent, and occasionally these curves actually crossed. Usually W_{RM} was greater than W_{CM}.

It should be noted that of the nine configurations considered, the LD_{50} was, at worst, located somewhat below the lowest dosage level.

The narrowest spacing was such that the probabilities ran from .59 to .86. A wider spacing was more favorable to the *SK* Method (as theory would indicate).

FIVE ANIMALS PER LEVEL

Since there are about thirteen hundred possible outcomes when there are four levels and five animals per level, a complete enumeration was beyond available resources. For the probability levels of Graph I, it was found that three hundred cases accounted for all but .1% of the probability, and Graph III indicates the results for these cases. No curve for the *CM* Method is shown as the calculation of this estimate for the three hundred cases was not feasible.

The *CM* estimate was calculated for nineteen outcomes constituting about 40% of the probability, but in most of these cases the correction was negligible. In the symmetrical situation, the true LD_{50} was midway between the second and the third level. When the original estimate was wide of the mark (i.e. toward the extreme levels), the maximum likelihood correction moved the estimate still further away from center in almost all cases. Hence the W_{CM} curve would be close to, but somewhat under, the W_{SK} curve.

This behavior of the maximum likelihood correction indicates that when the true value is located at or near the center of the levels, the *SK* solution will always be superior to the maximum likelihood solution. This result should hold regardless of m and n, although as the sample sizes increase, the difference becomes negligible.

The critical question therefore is, "What happens when the true value is not central?" Presumably the displacement produced by the iterative process will then be in the "right" direction and hence would do some good.

The asymmetrical probabilities (.59—.99) previously utilized in Graph II were also employed for the five animal case. It was found that 48 outcomes would account for slightly over 90% of the total probability, and the results of these outcomes are indicated in Graph IV.

For very small values of δ, W_{CM} is superior to W_{SK} although the difference is slight. For larger values of δ, W_{SK} gains a marked advantage. The reason for the failure of the maximum likelihood correction is that it "overcorrects"; although the *SK* estimate is an overestimate in most cases, the "corrected" estimate is an underestimate which is still further from the true value.

The curious twisting behavior of W_{CM} is due to the asymmetry of the distribution. If the sign of $T - \theta$ is taken into account, the curves are much more regular, although, of course, part of the irregularity is due to the omitted outcomes.

It should be noted that most of the probability in this asymmetric case is concentrated in outcomes which are such that a cautious research worker would probably prefer to repeat the experiment using different levels. Hence while it might be that for still more extreme asymmetry W_{CM} might gain the advantage, it would be of little practical importance, since it is not usual to attempt any estimate whenever the proportion dying at the lowest dosage is more than 50%.

Graph IV is unfair to the Reed-Muench Method in a sense because only *interpolative* estimates were used. Hence no *RM* estimate was possible in many of the cases. Extrapolative estimates were attempted and gave results rather similar to the *CM* estimates so that if such estimates were allowed, the W_{RM} curve would be higher.

CONCLUSIONS

While the small sample method is limited by the fact that only a few location-spacing configurations can be evaluated, the results of the cases investigated are consistent enough so that the following general conclusions may be stated with some confidence.

(1) The Spearman-Karber Method of estimating the LD_{50} provides superior estimates. It is also the simplest computationally.

(2) The Reed-Muench Method is fairly effective. It is also simple computationally. Some idea of the loss in accuracy entailed may be obtained by a study of Graphs I–IV.

(3) The Cornfield-Mantel iterative approximations to the maximum likelihood estimate do not improve the accuracy of the Spearman-Karber Method. The additional work seems to be wasted insofar as the LD_{50} is concerned.

It is highly interesting, although perhaps disconcerting, that older and computationally simpler methods should outperform the more "sophisticated" and difficult procedures derived by mathematical statisticians.

ACKNOWLEDGEMENTS

The small sample analysis was done under Navy contract N6onr-243. Computations made by Dr. Mario Pizzi for his doctoral dissertation were utilized in this analysis. Most of the remainder of the computation and the preparation of the graphs was done by Miss Marie M. Delaney. Miss Janise Harris and Miss Betty Grant helped on the iterative solutions. I am indebted to Professor William G. Cochran for valuable advice and criticism and to Dr. Joseph Berkson for several stimulating letters. The opinions expressed in this critique do not necessarily represent those of the aforementioned individuals.

REFERENCES

(1) Spearman, C., The Method of 'Right and Wrong Cases' (Constant Stimuli) without Gauss' Formulae, *Brit. Jour. of Psych., 2*, 1908.
(2) Irwin, J. O. and Cheeseman, E. N., On the Maximum Likelihood Method of Determining Dosage-Response Curves and Approximations to the Median-Effective Dose, in Cases of a Quantal Response, *Jour. Roy. Stat. Soc.* (Supplement), *6*, 1939.
(3) Finney, D. J., *Probit Analysis*, Cambridge, University Press, 1947.
(4) Cornfield, J. and Mantel, N., Simplified Calculation of the Dosage Response Curve, *Jour. Amer. Stat. Assoc., 45*, 1950.
(5) Reed, L. and Muench, H., A Simple Method of Estimating Fifty Percent Endpoints, *Amer. Jour. Hyg., 27*, 1938.
(6) Pizzi, M., An Approximate Solution to the Standard Error of the LD_{50} obtained by the Reed-Muench Method, Unpublished Ph. D. Dissertation, Johns Hopkins University, 1950.
(7) Berkson, J., Minimum X^2 and Maximum Likelihood Solution in Terms of Linear Transform, with Particular Reference to Bio-Assay, *Jour. Amer. Stat. Assoc., 44*, 1949.
(8) Thompson, W. R., Use of Moving Averages to Estimate Median-Effective Dose, *Bacteriological Review, 11*, 1947.

THE PLANNING OF PROBIT ASSAYS

M. J. R. Healy

Rothamsted Experimental Station

Summary—Graphs are presented from which the number of test subjects necessary to achieve a desired degree of precision in a 6-point probit assay may be determined.

The design of probit assays has been discussed by Finney (1947a) and the main considerations are well known. It is desirable for corresponding doses of the standard and unknown preparations to give approximately the same response, and the range of doses should not be such as to give responses near 100% or 0% as the weights attached to such responses are small. The most accurate design of assay is the 4-point in which two doses of each preparation are used, but as this does not give an adequate test of the assumptions of linearity and parallelism on which the assay is based the 6-point design is usually to be preferred. Three doses of each preparation are used, and these are best taken as equally spaced on the logarithmic scale.

In such an assay (assuming no control mortality or heterogeneity in the responses), the accuracy of the estimated relative potency can be roughly assessed in advance provided a preliminary estimate of the slope of the probit line can be made, as is often the case in practice. It is the purpose of this note to enable the user of probit assays to plan in advance the number of test subjects needed to attain the degree of precision which he requires. It is assumed that a 6-point design is used, and that equal numbers of test subjects are used at each dose.

Slopes of 1, 2, 4 and 8 have been taken, and for each slope six cases have been considered. First, the centre dose of the standard has been supposed to give a response of 5 probits (50%) or of 5.5 probits (about 70%). These situations will be referred to as *centred* and *uncentred* respectively. Secondly the logarithm of the relative potency (denoted by M) has been taken as 0, 0.15 and 0.30. For each case a number of spacings between doses have been investigated. A fairly wide spacing is necessary to determine the slope adequately, while too widely spaced doses produce undesirable extreme responses. For each situation there is thus an optimum spacing, which turns out to be almost independent

of the number of test subjects used. Recommended spacings in terms of the ratio between successive doses are given in the accompanying table.

RECOMMENDED DOSE-RATIOS

Slope \ M	0	0.15	0.30
1	1 : 5	1 : 8	1 : 10
2	1 : 2	1 : 3	1 : 5
4	1 : 1.6	1 : 2	1 : 2
8	1 : 1.2	1 : 1.4	1 : 1.4

The same ratios may be used for the centred and uncentred cases.

The 5% fiducial limits of the log relative potency are given by

$$M + \frac{g}{1-g}(M - \bar{x}_2 + \bar{x}_1)$$

$$\pm \frac{1.96}{b(1-g)} \sqrt{\left[(1-g)\left(\frac{1}{{}_1Snw} + \frac{1}{{}_2Snw}\right) + \frac{(\bar{x}_2 - \bar{x}_1 - M)^2}{\sum S_{xx}}\right]}$$

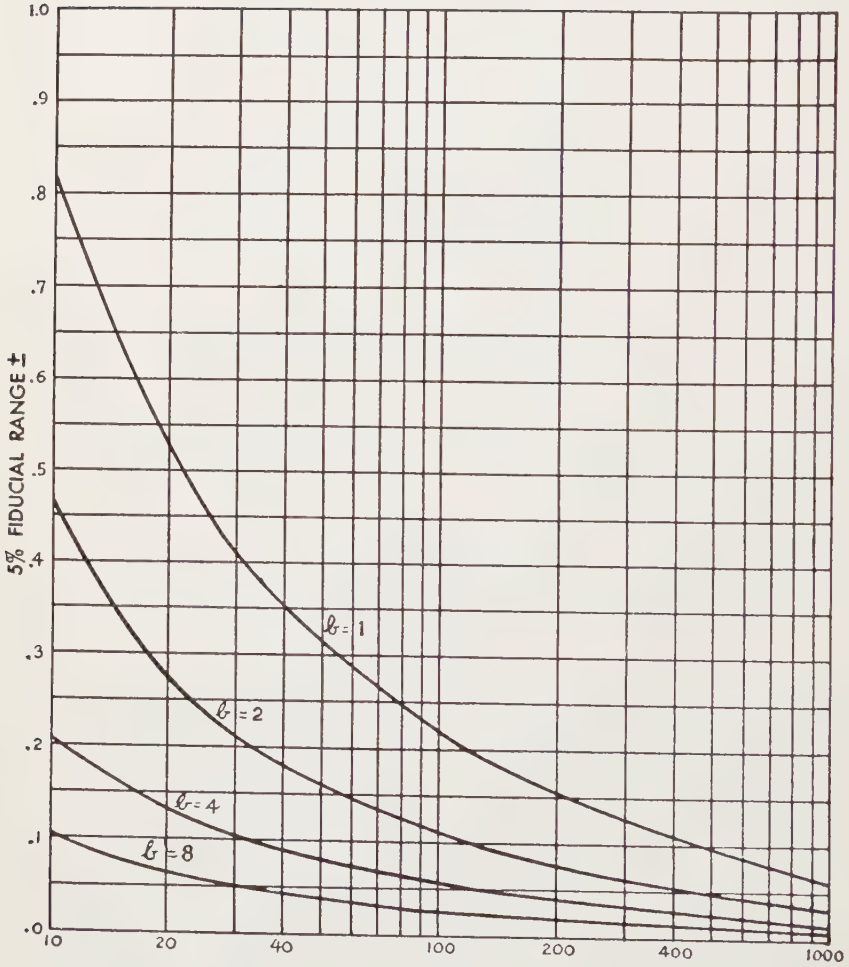
(cf. Finney 1947b, p. 66) where the symbols have their usual meanings. For each of the cases considered the recommended spacing was taken and the quantity following the ± sign was calculated for different values of n, the number of test subjects at each dose. The results are plotted in figs. 1–6.

The use of the graphs is best illustrated by an example. Suppose that an assay is being planned in which the standard preparation is known to give a slope of about 3, and that it is possible to arrange for a response close to 50% at the centre dose. Little is known about the unknown preparation, and it may be as much as twice as potent as the standard. It is desired to determine the log relative potency with 5% fiducial limits of, say, ±0.20. The situation corresponds to the centred case with $M = 0.30$. From the table it is found that successive doses should be in a ratio of 1:3.5, and Fig. 5 shows that about 30 test subjects should be used at each dose.

It should be noted that while the precision read from the graphs is that to be expected in the various cases, there is in general about a 50% chance of getting less accurate results. This is true of most of the com-

FIGURE 1

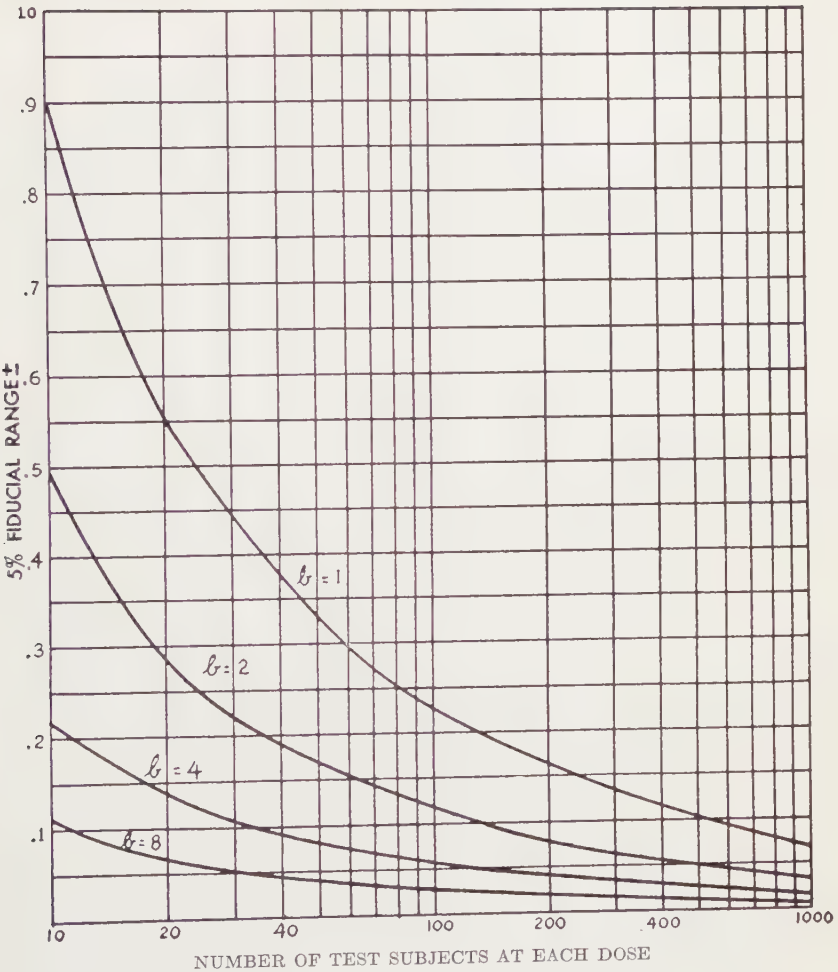
$M = 0$, CENTRED



monly used methods for predetermining the amount of material to be used in an experiment (Harris *et al.*, 1948). On the other hand, the results given for the uncentred cases are pessimistic, in that they assume that the effects of lack of centreing and of relative potency combine to reduce the expected precision, whereas these two factors may work in opposition.

FIGURE 2

$M = 0$, UNCENTRED



I am grateful to Dr. C. Potter and Dr. F. Yates for advice in the preparation of this paper.

REFERENCES

Finney, D. J. The principles of biological assay. *Suppl. J. Roy. Stat. Soc. 9*, 46, 1947a.

Finney, D. J. *Probit Analysis*. London: Cambridge University Press, 1947b.

Harris, M., Horvitz, D. G. and Mood, A. M. On the determination of sample sizes in designing experiments. *J. Amer. Stat. Ass., 43*, 391, 1948.

FIGURE 3

M = 0.15, CENTRED

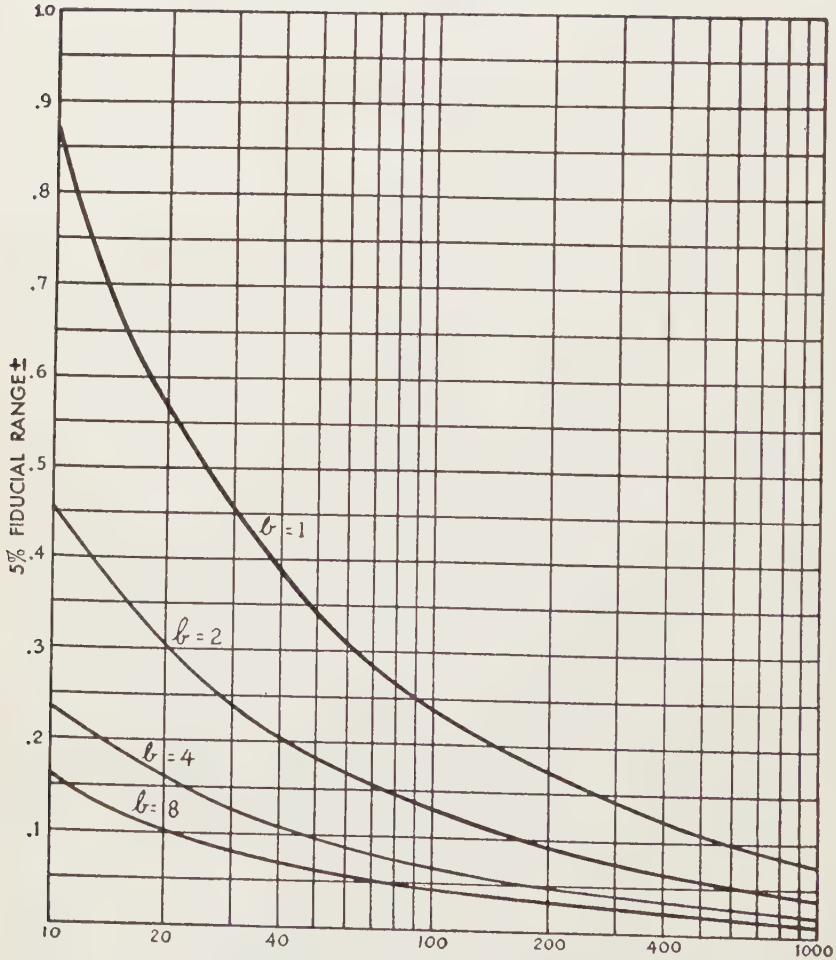


FIGURE 4

M = 0.15, UNCENTRED

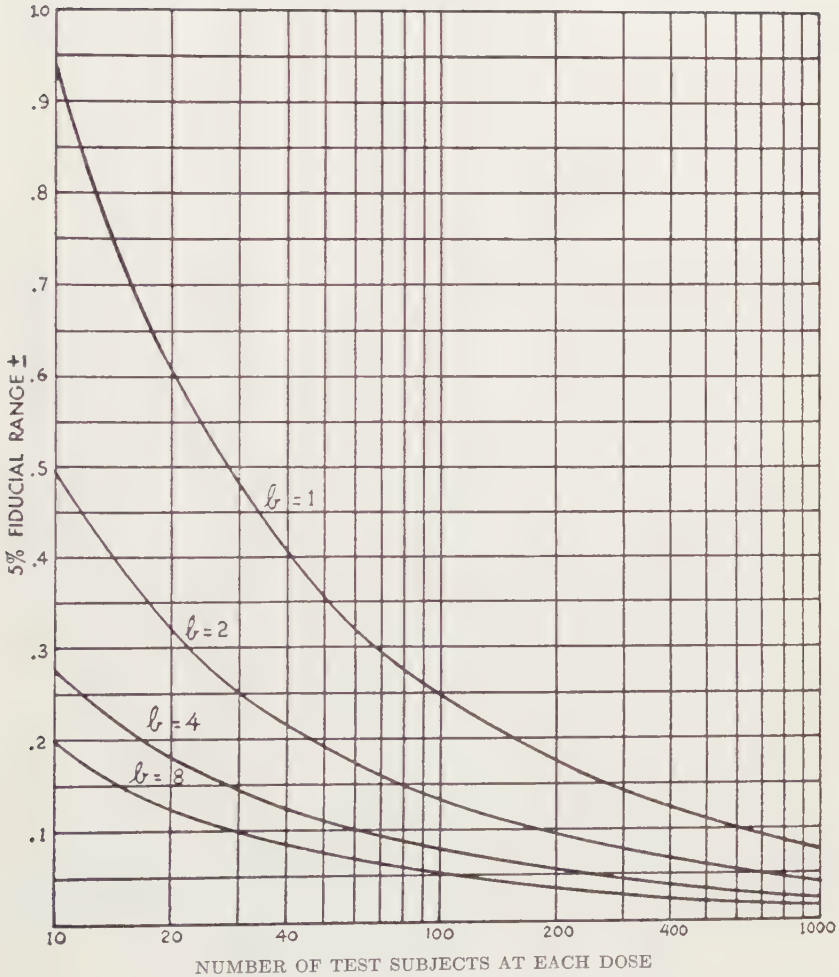


FIGURE 5

$M = 0.30$, CENTRED

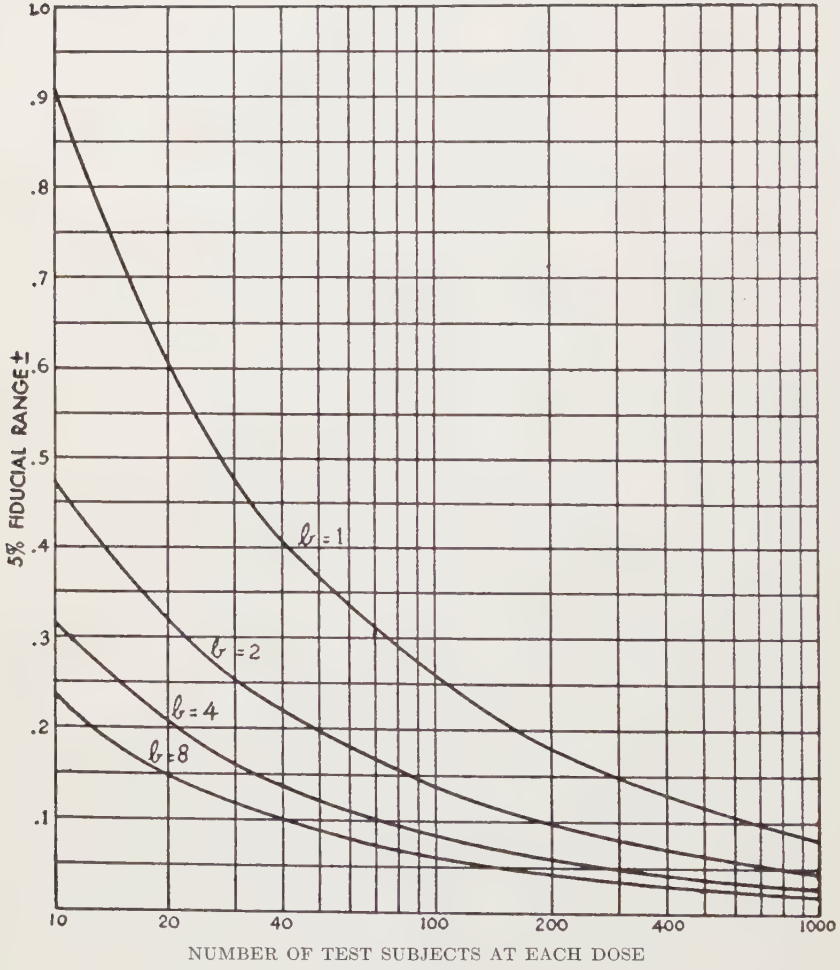
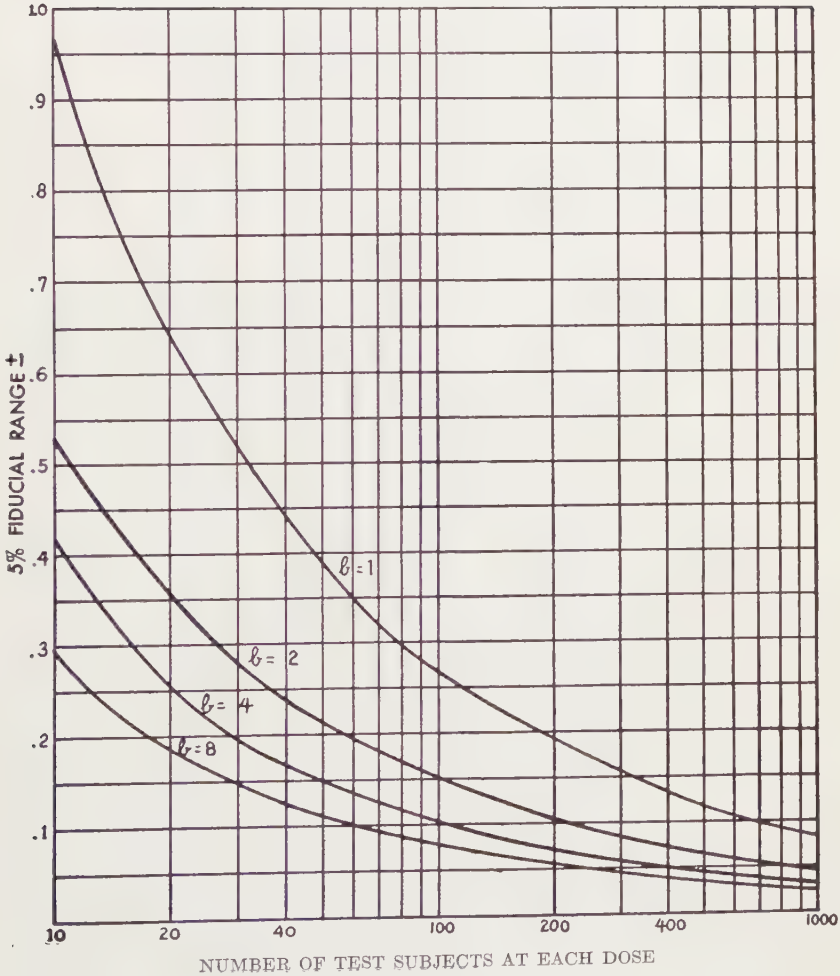


FIGURE 6

$M = 0.30$, UNCENTRED



SOME OBSERVATIONS WITH RESPECT TO THE ERROR OF BIO-ASSAY*

Joseph Berkson, M.D.

Division of Biometry and Medical Statistics,
Mayo Clinic,
Rochester, Minnesota

In table 1 is shown a comparison, for ten examples of bio-assay experiments obtained from published data, of the L.D. 50 as estimated by eight different methods: (1) probit, maximum likelihood; (2) logit, minimum X^2; (3) probit line fitted graphically; (4) logit line fitted graphically; (5) Reed-Muench method; (6) Kärber method; (7) Thompson moving average method; (8) straight line fitted by eye. The L.D. 50's are contained in the body of the table. In the bottom two rows are given the standard deviation of the estimates obtained by the different methods, and the ratio of this to the calculated standard deviation, the last obtained from the probit maximum likelihood solution by the usual formula.

The ratios varied from 0.13 to 0.68, seven of the ten being less than 0.5. Assuming the calculated standard error by the probit method to be the actual error of repeated experiments with the use of that method, this means that with a particular set of data before us and the potency estimated from the probit maximum likelihood line, if we were to compare the L.D. 50, on the one hand with the estimate obtained by another of the proposed methods, and on the other hand with an estimate obtained with the same probit method using another sample, the difference expected from the next sample would be twice as large as that from the alternative method. It is clear, therefore, in so far as these examples are representative of the data available for statistical bio-assay, that the *practical* importance of utilizing one rather than another of the proposed methods is generally small.

---

*The material in this note was presented in a paper delivered by the author at a joint session of the Biometrics Section of the American Statistical Association, and the Biometric Society, E.N.A.R., in New York, New York, December 30, 1949.

TABLE 1
L.D. 50 ESTIMATED BY DIFFERENT METHODS

Method	Example									
	1	2	3	4	5	6	7	8	9	10
1. Prbt. M.L.	0.686	0.995	1.23	0.982	0.939	0.907	0.092	0.983	0.903	0.199
2. Lgt. X^2	0.687	0.998	1.22	0.982	0.939	0.928	0.093	0.985	0.913	0.199
3. Prbt. graph	0.695	0.970	1.26	0.965	0.925	0.915	0.120	0.990	0.917	0.194
4. Lgt. graph	0.680	1.025	1.23	0.945	0.936	0.920	0.100	0.982	0.908	0.198
5. Reed-Muench	0.653	1.049	1.27	0.974	0.922	0.908	0.104	0.991	0.910	0.203
6. Kärber	—	—	—	—	0.940	0.916	0.090	0.985	0.916	—
7. Thompson, M.Av.	—	—	—	—	0.940	0.929	0.097	0.995	0.917	—
8. Arth. Log L.S.	0.689	1.014	1.24	0.994	0.976	0.947	0.105	1.019	0.921	0.203
σ meth.	0.015	0.027	0.020	0.017	0.016	0.013	0.010	0.012	0.006	0.003
σ meth./σ calc.	0.68	0.68	0.49	0.40	0.43	0.33	0.26	0.35	0.13	0.60

Some investigation was made to compare the standard error calculated by the usual formula with the standard error of estimates obtained by the probit maximum likelihood method, when experiments were repeated under essentially the same conditions. Only a small amount of data suitable for the investigation was available. It was found that the actual error was from three to ten times as large as the calculated error. These findings will have to be checked by the performance of an adequate number of experiments, properly designed for the purpose, before they can be considered definitive. But in so far as these preliminary results are representative, they indicate a serious defect in the statistical "model" utilized for the calculation of the error of the L.D. 50. The model utilized implies sampling from an existent population with no error in the independent variate, dosage. Actually the bio-assay experiment is a "controlled experiment" with error in dosage.[2] To clarify this question, mathematical investigation will have to be made of the statistics of experiments performed under the latter circumstance, as respects the error of estimate of the L.D. 50 and other aspects of procedures used, such as tests for linearity and tests for heterogeneity.

BIBLIOGRAPHY

1. Berkson, Joseph: Application of the Logistic Function in Bio-assay. *J. Am. Stat. Assoc.* 39:357–365, 1944.
2. Berkson, Joseph: Are There Two Regressions? *J. Am. Stat. Assoc.* 45: 164–180, 1950.
3. Bliss, C. I.: The Determination of the Dosage-mortality Curve From Small Numbers. *Quart. J. Pharm. & Pharmacol.* 11:192–216, 1938.
4. Finney, D. J.: *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve.* pp. 256, Cambridge, England, Cambridge University Press, 1947.
5. Reed, L. J. and Muench, H.: A Simple Method of Estimating Fifty Per Cent Endpoints. *Am. J. Hyg.* 27:493–497, 1938.
6. Thompson, W. R.: Use of Moving Averages and Interpolation to Estimate Median-effective Dose. *Bact. Rev.* 11:115–145, 1947.

QUERIES

85 QUERY: I serve as medico-legal advisor for the court. The law forbids people who are intoxicated with alcohol to drive cars, and drivers who have more than 0.05 percent alcohol in their blood are sentenced to jail. Whenever a car driver is suspected to be intoxicated, three blood samples from him are sent to my laboratory for analysis. My material looks like the following table.

Person	Percent Alcohol in Blood			
	Sample 1	Sample 2	Sample 3	Mean
1	0.08	0.06	0.06	0.067
2	0.14	0.18	0.16	0.160
3	0.00	0.02	0.01	0.010
		et cetera		

I have made determinations on the blood of several thousands of persons, and I want to make fiducial statements for my data, being willing to run only a minimal risk of being wrong, e.g. corresponding to the 0.999 999 probability point.

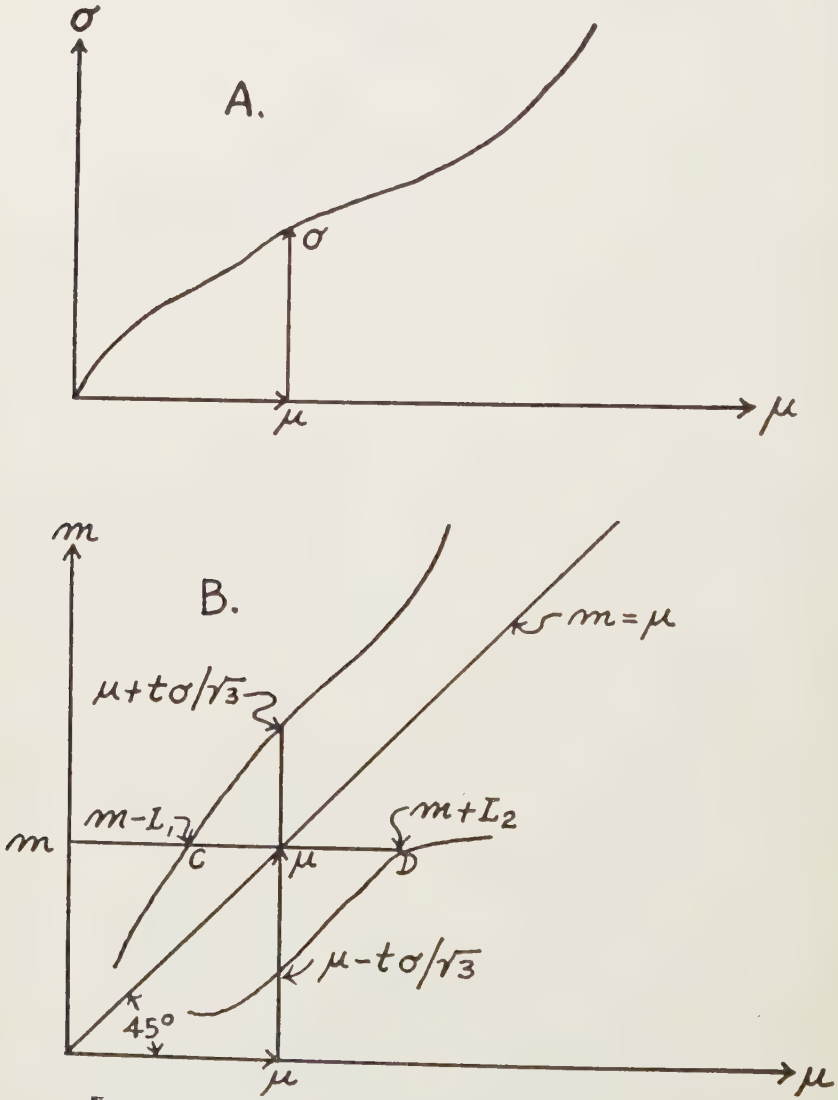
My practical problem is, based upon the whole material, to state fiducial limits for the alcohol concentration of each person, determined as the mean of 3 analyses, e.g. that person 1 has $0.067 \pm L$ percent alcohol in his blood.

ANSWER: Yours is a good idea, to calculate an average variance of the 3 samples per person, based upon the whole material. With the quantity of data you have, this average may be taken as specifying the standard deviation of the population of samples, subject to the following conditions.

First, your laboratory processes should be under some kind of statistical control so as to insure uniformity in your determinations. Historically, evidence can be got from examination of your standard deviations as a time series. We shall assume that there is neither trend nor short-time fluctuation in the standard deviations.

Second, any correlation there may be between mean and standard deviation must be eliminated or evaluated. A two-way table with means

m classed in intervals along one border and standard deviations s on the other should furnish the evidence. If m and s turn out to be independently distributed, then the mean of the variances s^2 may be taken as the population variance, σ^2. You will be assuming normal distribution of your samples from populations having common σ^2. The confidence interval for each person will be $m \pm t\sigma/\sqrt{3}$ where m is the mean of the 3 samples and t is a normal deviate selected at the desired probability level.



A. REGRESSION OF s ON m ASSUMED TO BE REGRESSION OF σ ON μ.
B. CONFIDENCE BELT FOR m.

If there is a regression of s on m two plans may be tried. If you can find a transformation that will eliminate the relation and will leave you with reasonably normal distributions, the transformed data can then be used for calculating σ^2 and for setting confidence intervals.

Incidentally, if you discover a regression then the time series of s should be separated into several series, one corresponding to each class of m.

If transformation is not successful, it would seem that the large quantity of your data will warrant the assumption that your regression of s on m adequately represents the regression of the corresponding population parameters, σ on μ. The curve may be like that in the upper part of the figure. Perhaps a graphical determination of the regression is sufficiently accurate, or you may wish to fit some function such as a polynomial.

For each μ, the probability is p that a mean of 3 observations will lie within the interval $\mu \pm t\sigma/\sqrt{3}$, t being again the chosen normal deviate. This interval is plotted on either side of the regression line $m = \mu$ in the lower part of the figure. The confidence belt between the curves may be used by inversion to specify CD, the confidence interval on μ, from $m - L_1$ to $m + L_2$, corresponding to the mean m of the percentage alcohol in a person's blood. It is clear that L_1 and L_2 may be expected to differ. These limits may be determined graphically or by use of the function fitted to the regression of s on m.

We have assumed that you have been calculating s for each person and that the results are available for study. If not, your large quantity of data warrants the use of the range of each trio of samples in lieu of s; for samples of 3, the efficiency of the range as compared to s is 0.99. In samples of 3 from populations with common σ, the average range multiplied by 0.591 is an estimate of σ.

Your suggestion of 0.999 999 as the probability desired seems to us unrealistic. In practice, distributions are never normal, whereas we have used the normal model in the foregoing suggestions. Again, we can only approximate σ and the regression of σ on μ. Also, there are doubtless changes in the accuracy of your laboratory techniques, making σ for any μ somewhat inconstant. These inaccuracies make it futile to seek probabilities such as the one specified. Our recommendation would be to select $t = 1.960$ at the 0.95 level or $t = 2.576$ for $P = 0.99$, then report to the Court the corresponding confidence interval with each mean. This information, along with other evidence, would be used by the Court to reach a decision.

C. I. Bliss
C. P. Winsor

THE BIOMETRIC SOCIETY

The Biometric Society was represented at the third World Health Assembly of the W.H.O. in Geneva last May by President Linder. At the sixth session of the W.H.O. Executive Board which followed, the official relations of the W.H.O. with non-governmental organizations were reviewed. In reporting its approval Dr. Brock Chisholm, Director General of W.H.O. wrote as follows: "The close and cordial relations which have been developed between your organization and W.H.O. in the field of mutual interests have already proved to be of great value to us, and I look forward with pleasure to an increase in their value in the future."

The Biometric Society serves as the Biometric Section of the International Union of Biological Sciences. Delegates of the Sections and officers of the IUBS met recently in Stockholm on July 7–11, primarily to review the activities of the Sections and to plan for the future. Professor Arthur Linder, President of the Society, and Dr. Bertil Matern, Statistician at the State Forest Research Institute in Stockholm, served as delegates of our Section.

Professor Linder summarized the activities of the Society and Section during the last three years and reviewed our tentative plans for the period 1950–1953. Financial support has been received or is being requested for several of these projects from the IUBS. One is the project on the teaching of biometry, which was initiated with a questionnaire sent to all members with their membership cards for 1950. A second is the publication of the proceedings of the Second International Biometric Conference in BIOMETRICS which was held last year in Geneva. The third project is an international symposium to be held in India in December, 1951, for which plans are now being initiated. This proposal was given first priority among IUBS-supported symposia for 1951, with the title "Problèmes biométriques dans la prévision et l'estimation de la croissance des végétaux dans les régions tropicales et subtropicales". The symposium will be synchronized with a meeting of the International Statistical Institute and with the annual meeting of the Indian Region of the Biometric Society.

Officers of the IUBS elected for the period 1950–1953 were President, Professor Munro Fox, Bedford College for Women, London; Vice-President, Professor J. Runnström, Stockholm. The session re-elected Secretary-General, Professor P. Vayssiere, Paris; Secretary, Professor Stuart Mudd, Philadelphia, and Treasurer, Professor F. Chodat, Geneva.

THE BIOMETRIC SOCIETY ACQUIRES BIOMETRICS

BIOMETRICS is now the journal of the Society, for which we are both financially and editorially responsible. It was started during 1945 as a small bimonthly "Biometrics Bulletin", primarily for the interchange of ideas in the absence of national meetings during the war. In 1947 it began quarterly publication, shortened its name to "Biometrics" and in December carried the proceedings of the international biometric conference in Woods Hole at which The Biometric Society was formed. The new Society arranged to have BIOMETRICS sent to all of its members who were not already subscribers.

As the Society grew, BIOMETRICS was continued as its official journal through the cooperation of the American Statistical Association and its Biometrics Section. It soon became apparent, however, that with our international status, we could not depend indefinitely upon the generosity of a national association. Early in 1949 the Society approached the ASA concerning the possibility of its transferring BIOMETRICS to the Society as an alternative to starting its own journal. Terms proposed for the transfer were discussed through 1949 and approved in principle at the annual meeting of the Biometrics Section in December and by the Board of Directors and Council of the ASA.

The ASA named Lowell J. Reed, its president-elect, and Harold F. Dorn, chairman of its Biometrics Section, to arrange the transfer with J. W. Hopkins, C. I. Bliss, Gertrude M. Cox and Arthur Linder, representing the Society. The agreement was put in legal form and signed in August 1950 as of March 1, 1950. Although The Biometric Society has assumed financial responsibility beginning with the first issue of this year, the September issue was the first to appear under our imprint.

The initial paragraph of the agreement notes that for the purpose of improving publication conditions in the field of biometry and with a view to the development of closer working relations between the Biometrics Section of the ASA and the Society, the two organizations have agreed to the transfer of BIOMETRICS by the ASA to the Society under terms which may be summarized as follows:

1. The ASA transfers to the Society its rights and title to the March 1950 and subsequent issues of BIOMETRICS but assumes all obligations in respect to the volume completed as of December 1949.

2. The Society assumes all obligations to produce the March 1950 and subsequent issues of BIOMETRICS.

3. The Society agrees that for five years from the beginning of 1950, all members of the ASA qualifying under Article 3, Section 1 of the ASA

constitution may order BIOMETRICS through the ASA at block subscription rates equal to the amount allocated to BIOMETRICS in the subscriptions of American members of the Society.

4. The ASA transfers to the Society all existing undistributed back issues of BIOMETRICS. For a period of five years, one-half of the net profits from the sale of such back issues shall revert to the ASA.

5. If the Society proposes to discontinue the publication of BIOMETRICS before the end of the last issue in 1954 it shall notify the ASA and upon receiving its written request shall transfer and reassign to ASA all its right and title to BIOMETRICS together with all undistributed back issues, free of any financial obligation.

6. In recognition of the support given to the journal in its initial years by the ASA and its Biometrics Section, the Society agrees that the title page of BIOMETRICS shall bear the words "Founded by the Biometrics Section of the American Statistical Association" or their equivalent.

7. The Society agrees that one member of the editorial board of BIOMETRICS may be nominated by the Biometrics Section of the ASA to serve for a term of three years from the date of such nomination, the nominee to be approved by the editor of BIOMETRICS.

8. This agreement shall inure to the benefit of and be binding upon the successors and assigns of the parties hereto.

Under Society sponsorship we hope that the Journal will continue its steady development and exert an increasingly broad and salutary influence upon biometrical standards all over the world. An editoral board of international character will shortly be named to determine policy. We have been fortunate in having Professor Gertrude M. Cox continue as editor, and in the support given by her associates in the Institute of Statistics of The University of North Carolina.

BIOMETRICS

Contents of Volume VI

INDEX